Federated Causal Inference in Heterogeneous Observational Data*

Ruoxuan Xiong† Allison Koenecke‡ Michael Powell§ Zhu Shen¶
Joshua T. Vogelstein‖ Susan Athey**

August 11, 2021

Abstract

Analyzing observational data from multiple sources can be useful for increasing statistical power to detect a treatment effect; however, practical constraints such as privacy considerations may restrict individual-level information sharing across data sets. This paper develops federated methods that only utilize summary-level information from heterogeneous data sets. Our federated methods provide doubly-robust point estimates of treatment effects as well as variance estimates. We derive the asymptotic distributions of our federated estimators, which are shown to be asymptotically equivalent to the corresponding estimators from the combined, individual-level data. We show that to achieve these properties, federated methods should be adjusted based on conditions such as whether models are correctly specified and stable across heterogeneous data sets.

Keywords: Causal Inference, Propensity Scores, Federated Learning, Multiple Data Sets

*This research is generously supported by Microsoft Research, the Office of Naval Research grant N00014-19-1-2468, and DARPA L2M program FA8650-18-2-7834. We are very grateful for many helpful comments and suggestions from Kristine Koutout and Molly Offer-Westort. Data for this project were accessed using the Stanford Center for Population Health Sciences Data Core. The PHS Data Core is supported by a National Institutes of Health National Center for Advancing Translational Science Clinical and Translational Science Award (UL1 TR001085) and from Internal Stanford funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

†Emory University, Department of Quantitative Theory and Methods, ruoxuan.xiong@emory.edu.

‡Microsoft Research New England, akoenecke@microsoft.com.

§Johns Hopkins University, Department of Biomedical Engineering, Institute for Computational Medicine, mpowel35@jhu.edu.

¶Stanford University, Department of Statistics, zhushen@stanford.edu.

‖Johns Hopkins University, Department of Biomedical Engineering, Institute for Computational Medicine, jovo@jhu.edu.

**Stanford University, Graduate School of Business, athey@stanford.edu.

1 Introduction

There are many settings where the same treatment is applied in different environments and where the data sets are stored separately for each environment. It would often be beneficial, if possible, to pool data across environments to estimate treatment effects (e.g., when the sample size in any one data set is too small to obtain precise estimates). However, there may be constraints preventing the combination of the data sets (e.g., legal constraints, privacy concerns, proprietary interests, or competitive barriers). Therefore, it is useful to develop analytical tools that can reap the benefits of data combination without actually pooling the data. Methods that accomplish this while sharing only summary-level information across data sets are referred to as "federated" learning methods. In this paper, we develop federated learning methods tailored to the problem of causal inference. The methods allow for heterogeneous treatment effects and outcome models across data sets, and we provide theoretical guarantees that our methods perform as well asymptotically as if the data sets were combined.

We apply these methods to a problem previously studied by Koenecke et al. (2021), who found evidence in two separate medical claims data sets that exposure to alpha blockers (a class of commonly prescribed drugs) reduced the risk of adverse outcomes for patients with acute respiratory distress. The two data sets are heterogeneous, with records from one data set reflecting more elderly patients and spanning a longer time horizon than the other data set. We apply the methods of this paper to combine the point and variance estimates from heterogeneous data sets, narrowing the confidence interval for these drugs' effects.

Multiple streams of literature study methods to analyze data from multiple sources, but most studies focus on pooling point estimates. However, studies of methods to pool confidence intervals are very limited and restricted to specific models. In machine learning, early developments provide methods to combine point estimates in linear models (Du et al., 2004, Karr et al., 2005), logistic models (Fienberg et al., 2006, Slavkovic et al., 2007), and maximum likelihood estimators (Blatt and Hero, 2004, Karr et al., 2007, Zhao and Nehorai, 2007, Lin and Karr, 2010) across distributed information systems, with most methods being iterative. Recent advances, mainly in federated learning, aim to develop communication-efficient methods to optimize the parameters in a more complex machine learning model such as deep neural networks, with a special focus on heterogeneous data sets and privacy considerations (Konečnỳ et al., 2016, McMahan et al., 2017, Li et al., 2020). Importantly, statistical inference is not a primary consideration in the aforementioned literature.

In biomedical studies, methods in meta-analysis and meta-regression analysis provide a weighted average from the results of the heterogeneous individual studies, where the weights take into account the uncertainties in the point estimates (e.g., inverse-variance weighting (DerSimonian and Laird, 1986, Whitehead and Whitehead, 1991, Sutton and Higgins, 2008, Hartung et al., 2011)). There is often only a single parameter of interest in meta-analysis and a linear functional form imposed in meta-regression analysis. Recent studies develop privacy-preserving methods to pool the summary-level information across multiple studies for broader classes of models such as linear models (Toh

et al., 2018, Li et al., 2019, Toh et al., 2020), logistic models (Li et al., 2016), Poisson models (Shu et al., 2019), and GLM (Wolfson et al., 2010)); only a few of these studies provide pooling methods for confidence intervals, but they are restricted to specific models and lack asymptotic theory for the pooling methods (e.g., Poisson models (Shu et al., 2019) or Cox models (Shu et al., 2020)). There has been a growing literature surrounding the development of causal inference methods for multiple data sets collected under heterogeneous conditions (e.g., Peters et al., 2016, Bareinboim and Pearl, 2016, Rosenman et al., 2018, 2020, Athey et al., 2020, Rothenhäusler et al., 2021). Vo et al. (2021) is most relevant to our paper and focuses on preserving data privacy by using Gaussian processes to model potential outcomes.

In this paper, we develop three categories of federated methods based on asymptotic theory that can combine the point and variance estimates from heterogeneous data sets. Our goal is to provide categories of methods that are applicable to a broad class of model specifications and are robust to outcome model misspecification.

Our first category of federated methods is for the point and variance estimates from MLE (which we refer to as "federated MLE"); this is used to develop the second and third categories of federated methods that achieve our aforementioned goal.[1] Our federated methods are based on the asymptotic theory of MLE (White, 1982). When the model is stable, our federated point estimate from MLE coincides with that from existing pooling methods for MLE (Blatt and Hero, 2004, Karr et al., 2007, Zhao and Nehorai, 2007, Lin and Karr, 2010). We additionally provide comparisons between our federated variance estimates and the prior literature to account for whether the model is misspecified. Moreover, when the model is unstable, our federated MLE is based on an adjusted pooled model for multiple data sets. Compared to the prior literature, we provide novel asymptotic theory for our federated MLE, which is required for the development of our second and third categories of federated methods.

Our second category of federated methods—which we believe to be particularly useful in empirical applications—is based on the Inverse Propensity-Weighted Maximum Likelihood Estimator (IPW-MLE), which uses the inverse of the estimated propensity score to construct weights for the likelihood function and to balance the covariate distribution across treatment and control groups. IPW-MLE meets our objectives and enjoys the doubly robust property in that it is consistent if we observe all relevant covariates and confounders, and either the propensity model or the outcome model is correctly specified (Bang and Robins, 2005, Wooldridge, 2007, Austin and Stuart, 2015).

Developing federated methods for IPW-MLE is underexplored in the literature and is challenging for three reasons. First, IPW-MLE involves both the propensity and outcome models, so the federated methods for IPW-MLE require theoretically guaranteed methods to federate both the propensity and outcome models. Second, many conditions can affect the federation of propensity and outcome models, such as whether the propensity model is known and whether the propensity and outcome models are correctly specified and stable (i.e., having a homogeneous response of outcome/treatment to confounders) across data sets. Third, if the propensity model is estimated, the

[1] Federated MLE on its own does not achieve our goal as it is not a doubly robust method.

federated method must take into account the propensity model's estimation error when estimating the outcome model from IPW-MLE (Wooldridge, 2002, 2007); however, this estimation error is often overlooked in practice, even in the setting of a single data set (e.g., in the standard SVYGLM package in R), leading to an overestimate of variance and a loss of efficiency.

We overcome these challenges and provide federated methods for IPW-MLE, which we refer to as "federated IPW-MLE." We focus on the case in which the propensity model is estimated from MLE, allowing us to use our federated MLE method for the propensity model. We then develop federated methods for the outcome model estimated from IPW-MLE; these methods are based on the asymptotic theory of IPW-MLE with a single data set (Wooldridge, 2002, 2007). Our federated IPW-MLE also varies with conditions such as whether the propensity and outcome models are correctly specified and stable across data sets.

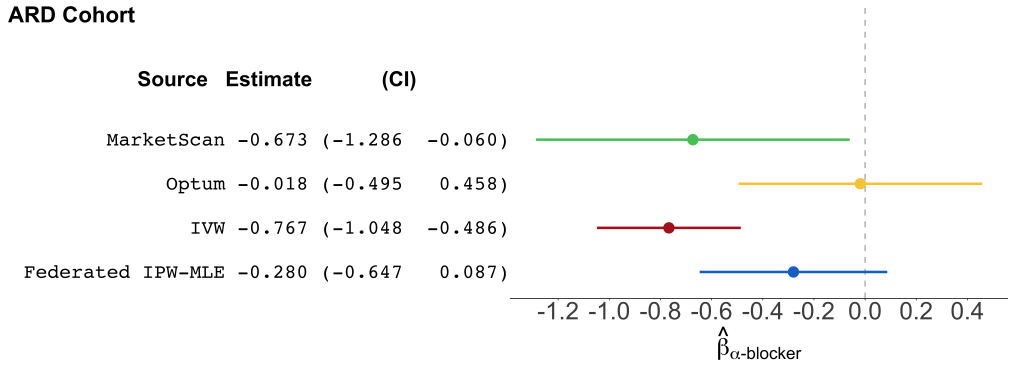
Let us revisit the example in Koenecke et al. (2021) to illustrate our federated IPW-MLE. We fit an IPW logistic regression model for each data set separately to estimate the effect of alpha blockers on preventing a patient suffering from acute respiratory distress from progressing to mechanical ventilation and then death. We consider two federated methods: one is our federated IPW-MLE, and the other is inverse variance weighting (IVW). IVW is commonly used in meta-analysis, and it is an appropriate method for (equally weighted) MLE assuming the outcome model is correctly specified and stable across data sets. Figure 1 shows the coefficient for alpha blocker exposure on two separate data sets as well as the federated coefficient from two federated methods. The federated coefficient from our federated IPW-MLE method lies between the coefficients estimated separately on two data sets, while the federated coefficient from IVW lies outside this interval, contradicting our intuition as we expect the federated coefficient to measure the average effect of alpha blockers across all patients in both data sets.[2] [3]

Our third category of federated methods is based on the Augmented Inverse Propensity Weighted (AIPW) Estimator (which we refer to as "federated AIPW"). AIPW also meets our objective and enjoys doubly robust properties (Robins et al., 1994, Kang et al., 2007, Tsiatis and Davidian, 2007). We focus on the case in which both the propensity and outcome models are estimated from MLE so that our federated MLE can be used as a building block to develop the federated methods for AIPW. We show that, similar to federated IPW-MLE, we need different federated methods for AIPW depending on whether the propensity and outcome models are stable across data sets. For empirical analyses, we focus on our federated IPW-MLE method over our federated AIPW method because their resulting metrics are not directly comparable—in particular, AIPW produces a measure based on ATE/ATT which is less intuitive to interpret.

[2]The main reason for the federated coefficient from IVW to lie outside this interval is that we have heterogeneous coefficients and variance-covariance matrices across datasets, which happens when different datasets have different populations. Consider a simplified example where we only have treatment and age in the outcome model. Assume we find that the age coefficient has opposite signs in the two data sets, and the covariance between estimated coefficients of treatment and age has opposite signs (which is the case on MS and Optum; see Figure 10 for age coefficients). Then, the federated treatment coefficient from IVW could lie outside the interval between treatment coefficients on two data sets. See Appendix B.1 for a numerical example.

[3]Similar analysis on an additional cohort of patients is provided in Figure 9 in Appendix B.

Figure 1: Coefficient of the Exposure to Alpha Blockers



Coefficient of the exposure to alpha blockers in an inverse propensity-weighted logistic regression to model the probability of progression to mechanical ventilation and then death of inpatients with acute respiratory distress. The top two lines show the estimated coefficient and 95% confidence interval on the IBM MarketScan®Research Database (which we refer to as MarketScan) and on Optum's Clinformatics®Data Mart Database (which we refer to as Optum). The third and fourth lines correspond to the federated coefficient and 95% confidence interval of MarketScan and Optum from inverse covariance-matrix weighting (i.e., IVW) and from our approach (i.e., unrestricted federated IPW-MLE), respectively. The federated coefficient from IVW contrasts with our expectation as its absolute value is larger than those on MarketScan and Optum. On the other hand, our approach provides a more reasonable pooled estimate, where the federated coefficient lies between those from MarketScan and Optum, and the confidence interval is narrower than those from MarketScan and Optum.

Our proposed methods improve upon the prior literature in two ways: first, the estimators are simple to use, and second, we have guarantees provided by asymptotic theory. Starting with ease of use, our federated estimators of the treatment effect and the associated variance do not require iteratively sharing summary-level information across data sets, which may be cumbersome or prohibitive; rather, the methods require only one-way, one-time sharing of summary-level information from each data set. In contrast, because there does not exist an explicit formula to estimate parameters in nonlinear models, prior work[4] suggests using an iterative approach based on the Newton-Raphson method for nonlinear models.[5] Second, we develop asymptotic theory for our federated point and variance estimators. All three categories of federated estimators (MLE, IPW-MLE, and AIPW) are efficient, consistent, and asymptotically normal. They all enjoy the optimal $n_{\text{pool}}^{1/2}$ convergence rate and have the same asymptotic distribution as the corresponding estimators using the combined, individual-level data, where n_{pool} is the number of observations on the combined data. Our federated variance estimators are consistent estimators for the asymptotic variances of the estimators on the pooled data. As such, we have proposed a feasible and simple approach to constructing valid confidence intervals for the pooled data without sharing individual-level information.

[4]For example, the federated estimators in logistic models (Fienberg et al., 2006, Slavkovic et al., 2007, Li et al., 2016), Poisson models (Shu et al., 2019), and MLE (Blatt and Hero, 2004, Karr et al., 2007, Zhao and Nehorai, 2007, Lin and Karr, 2010, Snoke et al., 2018).

[5]An iterative approach can provide estimators that are closer to those from the pooled individual-level data. However, we show that, asymptotically, the difference between iterative and non-iterative approaches is a higher order term that can be neglected. Therefore, we suggest a non-iterative approach.

The rest of the paper is organized as follows. Section 2 introduces the model and discusses the relevant covariate and model considerations/conditions for our federated estimators. Section 3 presents the details of our three federated point estimators and corresponding variance estimators under various conditions. The asymptotic distribution results for our federated estimators are provided in Section 4. In Section 5, we suggest a generalized federated analysis pipeline for empirical studies and apply it to study the two separate databases considered in Koenecke et al. (2021). We conclude in Section 6.

2 Model, Assumptions, and Preliminaries

In this section, we begin by stating the model setup and estimands for individual data sets in Section 2.1. Next, we review three widely used estimation approaches (MLE, IPW-MLE, and AIPW) on a single data set in Section 2.2. These three estimation approaches form the basis for our federated estimators, which will combine the summary-level information on individual data sets, and seek to obtain asymptotically equivalent estimates as these three approaches on the combined, individual-level data. Then, we list the covariate and model conditions that affect our federated estimators in Section 2.3. Finally, in Section 2.4, we state the three weighting methods to combine summary-level information that are used in our federated point and variance estimators. All of the matrices in the expression of the asymptotic variance for MLE and IPW-MLE are summarized in Table 1.

2.1 Model Setup

Suppose we have D data sets, where D is finite. For each data set $k \in \{1, \cdots, D\}$, we have n_k observations $(\mathbf{X}_i^{(k)}, Y_i^{(k)}, W_i^{(k)}) \in \mathcal{X}_k \times \mathbb{R} \times \{0, 1\}$ that are drawn i.i.d. from some distribution $\mathbb{P}^{(k)}$, and let $n_{\text{pool}} = \sum_{i=1}^{D} n_k$ be the total number of observations. Here, $i \in \{1, \cdots, n_k\}$ indexes the subjects (e.g., patients), $\mathbf{X}_i^{(k)}$ is a vector of d_k observed covariates, $Y_i^{(k)}$ is the outcome of interest, $W_i^{(k)}$ is the treatment assignment, and $\mathcal{X}_k \subseteq \mathbb{R}^{d_k}$. The number of covariates d_k can vary with data sets. Even if d_k is the same for all k, the covariates themselves could be different across data sets.

Under the Neyman-Rubin potential outcome model and the stable unit treatment value assumption (Imbens and Rubin, 2015), let $(Y_i^{(k)}(0), Y_i^{(k)}(1))$ be the outcome that subject i would have experienced if it had $(Y_i^{(k)}(1))$ or had not $(Y_i^{(k)}(0))$ been assigned treatment. For each data set k, suppose the standard unconfoundedness assumption (Rosenbaum and Rubin, 1983)

$$\{Y_i^{(k)}(0), Y_i^{(k)}(1)\} \perp W_i^{(k)} | \mathbf{X}_i^{(k)},$$

holds, and the overlap assumption (Rosenbaum and Rubin, 1983) for the propensity score $e^{(k)}(\mathbf{x}) = \text{pr}(W_i^{(k)} = 1 | \mathbf{X}_i^{(k)} = \mathbf{x})$,

$$\eta < e^{(k)}(\mathbf{x}) < 1 - \eta \quad \forall \mathbf{x} \in \mathcal{X}_k,$$

holds for some $\eta > 0$. For each data set k, we define the average treatment effect (ATE, denoted

as $\tau_{\text{ate}}^{(k)}$) and average treatment effect on the treated (ATT, denoted as $\tau_{\text{att}}^{(k)}$) as

$$\tau_{\text{ate}}^{(k)} \coloneqq \mathbb{E}[Y_i^{(k)}(1) - Y_i^{(k)}(0)], \quad \tau_{\text{att}}^{(k)} \coloneqq \mathbb{E}[Y_i^{(k)}(1) - Y_i^{(k)}(0) | W_i^{(k)} = 1]. \tag{1}$$

Our federated estimators seek to estimate the ATE and ATT on the combined, individual-level data.

2.1.1 Parametric Models

In this paper, we primarily focus on the cases where the outcome and propensity models are parametric as stated in Conditions 1 and 2 below. In medical applications, parametric outcome models are widely used (e.g., logistic regression for estimating the odds ratio in an epidemiological study (Sperandei, 2014), Cox regression for survival analysis in a clinical trial (Singh and Mukhopadhyay, 2011), and generalized linear models (GLM) for assessing medical costs (Blough et al., 1999, Blough and Ramsey, 2000)). To estimate the propensity model, one of the most common approaches is to use a logistic model (e.g., Imbens and Rubin (2015), Ch. 13). Moreover, the estimated parametric outcome and/or propensity model can then be used as the input in the estimation of ATE and ATT.

Condition 1 (Parametric Outcome Model)**.** *For any data set k, the conditional density function $f(y|\mathbf{x}, w)$ follows a parametric model, denoted as $f(y|\mathbf{x}, w, \boldsymbol{\beta})$ with parameters $\boldsymbol{\beta}$ and their true values $\boldsymbol{\beta}_0^{(k)}$.*

Condition 2 (Parametric Propensity Model)**.** *For any data set k, the conditional treatment probability $\mathbb{P}(w = 1|\mathbf{x})$ follows a parametric model, denoted as $e(\mathbf{x}, \boldsymbol{\gamma})$, with parameters $\boldsymbol{\gamma}$ and their true values $\boldsymbol{\gamma}_0^{(k)}$.*

We can estimate the parameters in the parametric outcome or propensity model by maximizing the (weighted) likelihood function that is presented for completeness in Section 2.2. Our federated estimators seek to estimate parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in the outcome and propensity models on the combined, individual-level data by combining estimates on individual data sets.

The following assumption is necessary to show the MLE is consistent and asymptotically normal (e.g., Newey and McFadden (1994)).

Assumption 1 (Parametric Outcome and Propensity Models)**.**

1. *Condition 1 holds. For any k, $\mathcal{X}_k$ is bounded. $f(y|\mathbf{x}, w, \boldsymbol{\beta})$ is twice continuously differentiable. $\boldsymbol{\beta}_0^{(k)} \in \mathcal{S}_{\boldsymbol{\beta}}^{(k)} \subset \mathbb{R}^{d_k+1}$ lies in the interior of a known compact set $\mathcal{S}_{\boldsymbol{\beta}}^{(k)}$. The information matrix $\mathcal{I}^{(k)}(\boldsymbol{\beta}) = -\mathbb{E}_{(\mathbf{x},w,y)\sim\mathbb{P}^{(k)}}\left[\frac{\partial^2 \log f(y|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right]$ is positive definite, is of full rank, and its condition number is bounded for all $\boldsymbol{\beta}$.*

2. *Condition 2 holds. For any k, $\mathcal{X}_k$ is bounded. $e(\mathbf{x}, \boldsymbol{\gamma})$ is twice continuously differentiable. $\boldsymbol{\gamma}_0^{(k)} \in \mathcal{S}_{\boldsymbol{\gamma}}^{(k)} \subset \mathbb{R}^{d_k+1}$ lies in the interior of a known compact set $\mathcal{S}_{\boldsymbol{\gamma}}^{(k)}$. The information matrix $\mathcal{I}^{(k)}(\boldsymbol{\gamma}) = -\mathbb{E}_{(\mathbf{x},w)\sim\mathbb{P}^{(k)}}\left[\frac{\partial^2 \log e(\mathbf{x},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top}\right]$ is positive definite, is of full rank, and its condition number is bounded for all $\boldsymbol{\gamma}$.*

2.2 Estimation on a Single Data Set

In this subsection, we review three estimation approaches (MLE, IPW-MLE, and AIPW) on a single data set. The estimates from these three approaches on the combined, individual-level data, which can be viewed as a single data set, are the values that our federated MLE, IPW-MLE, and AIPW estimators seek to produce.

2.2.1 Maximum Likelihood Estimation

Under the parametric outcome model (Condition 1), the conditional likelihood function of the outcome on each data set can be factorized, since observations are independently and identically distributed. That is,

$$\mathcal{L}_n(\boldsymbol{\beta}) = f(y_1, \cdots, y_n | \mathbf{x}_1, \cdots, \mathbf{x}_n, w_1, \cdots, w_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i | \mathbf{x}_i, w_i, \boldsymbol{\beta}).$$

The corresponding log-likelihood function with the natural logarithm is

$$\boldsymbol{\ell}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i | \mathbf{x}_i, w_i, \boldsymbol{\beta}). \tag{2}$$

The estimator $\hat{\boldsymbol{\beta}}_{\text{mle}}$ maximizes the log-likelihood function. That is,

$$\hat{\boldsymbol{\beta}}_{\text{mle}} = \arg\max_{\boldsymbol{\beta}} \boldsymbol{\ell}_n(\boldsymbol{\beta}). \tag{3}$$

Under the parametric propensity model (Condition 2), we can analogously estimate the parameters $\boldsymbol{\gamma}$ in the propensity model by MLE.

2.2.2 Inverse Propensity-Weighted Maximum Likelihood Estimation

Under Condition 1, an alternative approach to estimating $\boldsymbol{\beta}$ in the outcome model is to use the Inverse Propensity-Weighted Maximum Likelihood estimator (IPW-MLE, or pseudo/weighted maximum likelihood estimator). Inverse propensity weighting dates back to Horvitz and Thompson (1952) to estimate the population mean when data is nonrandomly missing, and has been extensively studied and used thereafter to address a variety of sample selection problems (e.g., Robins et al. (1995), Robins and Rotnitzky (1995), Hirano et al. (2003)), including those inherent in estimating ATE and ATT. IPW-MLE is motivated by doubly robust estimators in causal inference (Bang and Robins, 2005, Kang et al., 2007) using the following objective function with the natural logarithm:

$$\boldsymbol{\ell}_n(\boldsymbol{\beta}, \hat{e}) = \sum_{i=1}^{n} \varpi_{i,\hat{e}} \log f(y_i | \mathbf{x}_i, w_i, \boldsymbol{\beta}), \tag{4}$$

where $\varpi_{i,\hat{e}}$ is the weight for unit i and $\hat{e}$ is the abbreviation of $\hat{e}(\mathbf{x}_i)$ (i.e., the estimated $e(\mathbf{x}_i)$). For example, if $\varpi_{i,\hat{e}} = \frac{w_i}{\hat{e}(\mathbf{x}_i)} + \frac{1-w_i}{1-\hat{e}(\mathbf{x}_i)}$, then $\varpi_{i,\hat{e}}$ is the ATE weight; if $\varpi_{i,\hat{e}} = w_i + \frac{\hat{e}(\mathbf{x}_i)}{1-\hat{e}(\mathbf{x}_i)}(1-w_i)$, then $\varpi_{i,\hat{e}}$ is the ATT weight. Let $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ be the solution that maximizes the weighted log-likelihood function (4),

$$\hat{\boldsymbol{\beta}}_{\text{ipw-mle}} = \arg\max_{\boldsymbol{\beta}} \boldsymbol{\ell}_n(\boldsymbol{\beta}, \hat{e}). \tag{5}$$

$\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ has a general double robustness property in that $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ is consistent if (a) we have observed all relevant covariates and (b) at least one of the propensity and outcome models is correctly specified (Wooldridge, 2007, Lumley, 2011). When we are worried about the outcome model misspecification, IPW-MLE could be preferable to MLE as MLE could be inconsistent. Not surprisingly, the double robustness property comes at a cost: $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ could be less efficient than $\hat{\boldsymbol{\beta}}_{\text{mle}}$ if the outcome model is correctly specified.

2.2.3 AIPW Estimation

To estimate τ_{ate} or τ_{att} defined in Eq. (1), we can use the augmented inverse-propensity weighted (AIPW) estimator that takes the form

$$\hat{\tau}_{\text{aipw}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(\mathbf{X}_i, W_i, Y_i), \tag{6}$$

where $\hat{\phi}(\mathbf{X}_i, W_i, Y_i)$ is the score. If the estimand is τ_{ate},

$$\hat{\phi}(\mathbf{X}_i, W_i, Y_i) = \hat{\mu}_{(1)}(\mathbf{X}_i) - \hat{\mu}_{(0)}(\mathbf{X}_i) + \frac{W_i}{\hat{e}(\mathbf{X}_i)}\big(Y_i - \hat{\mu}_{(1)}(\mathbf{X}_i)\big) - \frac{(1-W_i)}{1-\hat{e}(\mathbf{X}_i)}\big(Y_i - \hat{\mu}_{(0)}(\mathbf{X}_i)\big), \tag{7}$$

where $\mu_{(w)}(\mathbf{X}_i) = \mathbb{E}[Y_i | X_i, W_i = w]$ and $\hat{\mu}_{(w)}(\mathbf{X}_i)$ is the estimated $\mu_{(w)}(\mathbf{X}_i)$. If the estimand is τ_{att},

$$\hat{\phi}(\mathbf{X}_i, W_i, Y_i) = W_i\big(Y_i - \hat{\mu}_{(1)}(\mathbf{X}_i)\big) - \frac{\hat{e}(\mathbf{X}_i)(1-W_i)}{1-\hat{e}(\mathbf{X}_i)}\big(Y_i - \hat{\mu}_{(0)}(\mathbf{X}_i)\big). \tag{8}$$

AIPW has two prominent properties. First, similar to IPW-MLE, AIPW is also doubly robust (Robins et al., 1994, Scharfstein et al., 1999, Bang and Robins, 2005, Kang et al., 2007, Tsiatis and Davidian, 2007). Second, AIPW is semiparametric efficient, and its asymptotic variance depends on whether the propensity and outcome models are correctly specified. Lemma 2 in Section 4.3 provides the expressions for asymptotic variance under various cases.

2.3 Covariate and Model Considerations in Federated Estimators

Our federated point and variance estimators are based on the three estimators (MLE, IPW-MLE, AIPW) described in Section 2.2 and aim to obtain asymptotically equivalent estimates to the corresponding estimates on the combined, individual-level data. Our federated estimators combine

the summary-level information on individual data sets. To achieve the asymptotic equivalence properties, the methods for federation need to vary based on the multiple conditions listed below.

Condition 3 (Stable Covariate Distribution)**.** *The set of covariates and their joint distribution are the same across all data sets. That is,* $d_j = d_k$ *and* $\mathbb{P}^{(j)}(\mathbf{x}) = \mathbb{P}^{(k)}(\mathbf{x})$ *for any two data sets* j *and* k*.*

Condition 4 (Known Propensity Score)**.** *For all data sets, the true propensity scores are known and used to estimate parameters in the outcome model.*

Condition 5 (Stable Propensity Model)**.** *For all data sets, the set of covariates and their associated parameters in the propensity model are the same. That is,* $\boldsymbol{\gamma}_0^{(j)} = \boldsymbol{\gamma}_0^{(k)}$ *for any two data sets* j *and* k*.*

Condition 6 (Stable Outcome Model)**.** *For all data sets, the set of covariates and their associated parameters in the outcome model are the same. That is,* $\boldsymbol{\beta}_0^{(j)} = \boldsymbol{\beta}_0^{(k)}$ *for any two data sets* j *and* k*.*

Condition 7 (Correct Propensity Model Specification)**.** *For all data sets, the propensity model is correctly specified in the estimation. There are no omitted covariates, and the functional form of each covariate is correct.*

Condition 8 (Correct Outcome Model Specification)**.** *For all data sets, the outcome model is correctly specified in the estimation. There are no omitted covariates, and the functional form of each covariate is correct.*

We refer to heterogeneous data sets as the case where Condition 3 is violated. As a preview of our federated estimators provided in Section 3, our federated point estimators depend only on whether Conditions 4, 5, and 6 hold, while our federated variance estimators depend on Conditions 4 through 8. Our federated estimators work for both homogeneous and heterogeneous data sets (whether Condition 3 holds or not), and Condition 3 alone does not directly affect what our federated point and variance estimators look like. However, Conditions 5 and 6 are more likely to be violated when Condition 3 is violated, resulting in different federated estimators. Moreover, when Condition 3 is violated, we require stronger technical conditions to provide theoretical guarantees for our federated estimators. For example, we require that the relative sizes of the data sets grow at the same rate, as stated in Assumption 2 below, so that the asymptotic variances of the estimators on the pooled data are well defined.

Assumption 2 (Sample Size)**.** *For any data set* k*, there exists a* p_k *bounded away from 0 and 1, such that* $\lim_{n_1,\cdots,n_D\to\infty} \frac{n_k}{n_{\text{pool}}} = p_k$*, where* $n_{\text{pool}} = \sum_{k=1}^{D} n_k$*.*

2.4 Three Weighting Methods

Our federated point and variance estimators combine summary-level information across data sets using the following three weighting methods.

2.4.1 Hessian Weighting

Hessian weighting is used in the federated point estimators of the parameters in the parametric outcome and/or propensity model. For example, for the coefficients of the outcome model, we refer to Hessian weighting as

$$\hat{\boldsymbol{\beta}}^{\text{fed}} = \Big(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} \Big)^{-1} \Big(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} \hat{\boldsymbol{\beta}}^{(k)} \Big), \quad \text{where } \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} = \frac{\partial^2 \boldsymbol{\ell}_{n_k}(\hat{\boldsymbol{\beta}}^{(k)})}{\partial \boldsymbol{\beta}^{(k)} (\partial \boldsymbol{\beta}^{(k)})^\top}. \tag{9}$$

For the coefficients of the propensity model, we replace $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}$ by $\hat{\boldsymbol{\gamma}}^{(k)}$ and $\hat{\mathbf{H}}_{\boldsymbol{\gamma}}^{(k)}$, respectively, in Eq. (9).

2.4.2 Sample Size Weighting

Sample size weighting is used in the federated point estimators of τ_{ate} and τ_{att}, and almost all of the federated variance estimators (see more details in Tables 2-4). For some generic matrix $\mathbf{M}$, we refer to sample size weighting as

$$\mathbf{M}^{\text{fed}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \mathbf{M}^{(k)}, \quad \text{where } n_{\text{pool}} = \sum_{k=1}^{D} n_k. \tag{10}$$

2.4.3 Inverse Variance Weighting

Inverse variance weighting (IVW) is used in the federated point and variance estimators of τ_{ate} and τ_{att}. For some generic point estimator $\hat{\boldsymbol{\nu}}$ (which could be either a scalar or vector), we refer to inverse variance weighting as

$$\hat{\boldsymbol{\nu}}^{\text{fed}} = \Big(\sum_{k=1}^{D} (\hat{\mathbf{V}}_{\boldsymbol{\nu}}^{(k)})^{-1} \Big)^{-1} \Big(\sum_{k=1}^{D} (\hat{\mathbf{V}}_{\boldsymbol{\nu}}^{(k)})^{-1} \boldsymbol{\nu}^{(k)} \Big), \quad \text{where } \hat{\mathbf{V}}_{\boldsymbol{\nu}}^{(k)} = \widehat{\text{Var}}(\hat{\boldsymbol{\nu}}^{(k)}) \tag{11}$$

$$\hat{\mathbf{V}}_{\boldsymbol{\nu}}^{\text{fed}} = n_{\text{pool}} \Big(\sum_{k=1}^{D} (\hat{\mathbf{V}}_{\boldsymbol{\nu}}^{(k)})^{-1} \Big)^{-1}. \tag{12}$$

Note that $\hat{\mathbf{V}}_{\boldsymbol{\nu}}^{(k)}$ scales with $1/n_k$, consistent with the definition of variance in inverse variance weighting in meta analysis (e.g., DerSimonian and Laird (1986), Hartung et al. (2011)). On the other hand, $\hat{\mathbf{V}}_{\boldsymbol{\nu}}^{\text{fed}}$ does not scale with the sample size n_{pool}, which simplifies the notation in the presentation of the asymptotic results of our federated estimators in Section 4 (e.g., the consistency of $\hat{\mathbf{V}}_{\boldsymbol{\nu}}^{\text{fed}}$).

Table 1: A Summary of Matrices in the Asymptotic Variance of MLE and IPW-MLE

Matrix	Expression	Matrix	Expression
$\mathbf{A}_{\boldsymbol{\beta}}$	$\mathbb{E}\Big[-\frac{\partial^2 \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\Big]$	$\mathbf{A}_{\boldsymbol{\gamma}}$	$\mathbb{E}\Big[-\frac{\partial^2 \log e(\mathbf{x},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top}\Big]$
$\mathbf{B}_{\boldsymbol{\beta}}$	$\mathbb{E}\Big[\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big(\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big)^\top\Big]$	$\mathbf{B}_{\boldsymbol{\gamma}}$	$\mathbb{E}\Big[\frac{\partial \log e(\mathbf{x},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\big(\frac{\partial \log e(\mathbf{x},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\big)^\top\Big]$
ATE weighting $\varpi_{i,e_{\boldsymbol{\gamma}}} = \frac{w_i}{e_{\boldsymbol{\gamma}}(\mathbf{x}_i)} + \frac{1-w_i}{1-e_{\boldsymbol{\gamma}}(\mathbf{x}_i)}$		ATT weighting $\varpi_{i,e_{\boldsymbol{\gamma}}} = w_i + \frac{e_{\boldsymbol{\gamma}}(\mathbf{x}_i)}{1-e_{\boldsymbol{\gamma}}(\mathbf{x}_i)}(1-w_i)$	
$\mathbf{A}_{\boldsymbol{\beta},\varpi}$	$\mathbb{E}\Big[\big(\frac{w}{e_{\boldsymbol{\gamma}}} + \frac{1-w}{1-e_{\boldsymbol{\gamma}}}\big)\frac{\partial^2 \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\Big]$	$\mathbf{A}_{\boldsymbol{\beta},\varpi}$	$\mathbb{E}\Big[\big(w + \frac{e_{\boldsymbol{\gamma}}(1-w)}{1-e_{\boldsymbol{\gamma}}}\big)\frac{\partial^2 \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\Big]$
$\mathbf{D}_{\boldsymbol{\beta},\varpi}$	$\mathbb{E}\Big[\big(\frac{w}{e_{\boldsymbol{\gamma}}} + \frac{1-w}{1-e_{\boldsymbol{\gamma}}}\big)^2\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\cdot\big(\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big)^\top\Big]$	$\mathbf{D}_{\boldsymbol{\beta},\varpi}$	$\mathbb{E}\Big[\big(w + \frac{e_{\boldsymbol{\gamma}}(1-w)}{1-e_{\boldsymbol{\gamma}}}\big)^2\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\cdot\big(\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\big)^\top\Big]$
$\mathbf{C}_{\boldsymbol{\beta},\varpi}$	$\mathbb{E}\Big[\big(\frac{w}{e_{\boldsymbol{\gamma}}^2} - \frac{1-w}{(1-e_{\boldsymbol{\gamma}})^2}\big)\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\cdot\big(\frac{\partial \log e(\mathbf{x},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\big)^\top\Big]$	$\mathbf{C}_{\boldsymbol{\beta},\varpi,1}$	$\mathbb{E}\Big[-\frac{(1-w)}{(1-e_{\boldsymbol{\gamma}})^2}\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\cdot\big(\frac{\partial \log e(\mathbf{x},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\big)^\top\Big]$
		$\mathbf{C}_{\boldsymbol{\beta},\varpi,2}$	$\mathbb{E}\Big[\big(\frac{w}{e_{\boldsymbol{\gamma}}} - \frac{e_{\boldsymbol{\gamma}}(1-w)}{(1-e_{\boldsymbol{\gamma}})^2}\big)\frac{\partial \log f(y\|\mathbf{x},w,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\cdot\big(\frac{\partial \log e(\mathbf{x},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\big)^\top\Big]$

In the definitions of these matrices, $e_{\boldsymbol{\gamma}}$ denotes $e_{\boldsymbol{\gamma}}(\mathbf{x}_i) = e(\mathbf{x}_i, \boldsymbol{\gamma})$ by a slight abuse of notation.

3 Three Federated Estimators

We refer to the federated estimators based on MLE, IPW-MLE, and AIPW as federated MLE, federated IPW-MLE, and federated AIPW, respectively. Our main focus is the doubly robust federated methods—IPW-MLE and AIPW—but we first introduce the federated MLE on which federated IPW-MLE and AIPW are built. For each category of federated estimator (federated MLE, IPW-MLE, or AIPW), we start with the simple case in which the propensity and outcome models are stable (Conditions 5 and 6 hold). In this case, it is natural to use the model for the combined, individual-level data that has the same functional form and model parameters as those for each individual data set, since the functional form and model parameters are the same across individual data sets. We refer to our federated estimators in this case as **restricted** federated estimators. Our restricted federated estimators leverage the invariant functional form and parameters and, as will be shown in Section 4, they are efficient and asymptotically equivalent to the corresponding estimators from the combined, individual-level data.

Next, we generalize our restricted federated point and variance estimators to the more challenging case in which at least one of the propensity and outcome models is unstable (either Condition 5 or 6 is violated). Our federated estimators in this case are referred to as **unrestricted** federated estimators, which are built on the corresponding restricted federated estimators, and allow for parameters and their values in the propensity or outcome models to be specific to individual data sets.

Flowcharts to represent how our federated point and variance estimators vary with various conditions are provided in Figures 2-6, with more details about our federated estimators provided in Tables 2-4. The theoretical guarantees of our federated estimators are deferred to Section 4.

3.1 Federated Maximum Likelihood Estimator

3.1.1 Restricted Estimator for Stable Models (Conditions 5 and 6 hold)

Our federated MLE works for both parametric outcome (Condition 1) and parametric propensity (Condition 2) models. In this subsection, we illustrate how our federated MLE works using the outcome model. Our federated MLE works in the same fashion for the propensity model.

For the restricted federated point estimator, we first estimate parameters $\hat{\boldsymbol{\beta}}^{(k)}_{\text{mle}}$ using MLE for each individual data set k and then use Hessian weighting (9) to combine $\hat{\boldsymbol{\beta}}^{(k)}_{\text{mle}}$ across all data sets. This procedure is the same regardless of whether the covariate distribution is stable (Condition 3) or whether the outcome model is correctly specified (Condition 8), as shown in Table 2.

Our proposed restricted federated variance estimator is based on the expression of the asymptotic variance of a single data set since we seek to obtain asymptotically equivalent variance as that from the combined, individual-level data. The asymptotic variance of a single data set, such as the combined, individual-level data, varies with whether the (outcome) model is correctly specified (White, 1982). If the model is misspecified, the asymptotic variance is $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta}}^{-1}\mathbf{B}_{\boldsymbol{\beta}}\mathbf{A}_{\boldsymbol{\beta}}^{-1}$ (also known as the "sandwich formula"), where the definition of $\mathbf{A}_{\boldsymbol{\beta}}$ and $\mathbf{B}_{\boldsymbol{\beta}}$ can be found in Table 1. On the other hand, if the model is correctly specified, then the information matrix equivalence holds (i.e., $\mathbf{B}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta}}$), and the asymptotic variance is simplified to $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta}}^{-1}$ (referred to as the "simplified formula"). The federated variance estimator consists of three steps: first, estimate $\mathbf{A}^{(k)}_{\boldsymbol{\beta}}$ (and $\mathbf{B}^{(k)}_{\boldsymbol{\beta}}$ if necessary) on each individual data set k; second, combine $\mathbf{A}^{(k)}_{\boldsymbol{\beta}}$ (and $\mathbf{B}^{(k)}_{\boldsymbol{\beta}}$ if necessary) across all data sets via sample size pooling (10) to estimate the pooled $\mathbf{A}^{\text{fed}}_{\boldsymbol{\beta}}$ (and $\mathbf{B}^{\text{fed}}_{\boldsymbol{\beta}}$ if necessary); third, return $(\mathbf{A}^{\text{fed}}_{\boldsymbol{\beta}})^{-1}$ as the pooled variance if the model is correctly specified and $(\mathbf{A}^{\text{fed}}_{\boldsymbol{\beta}})^{-1}\mathbf{B}^{\text{fed}}_{\boldsymbol{\beta}}(\mathbf{A}^{\text{fed}}_{\boldsymbol{\beta}})^{-1}$ otherwise.

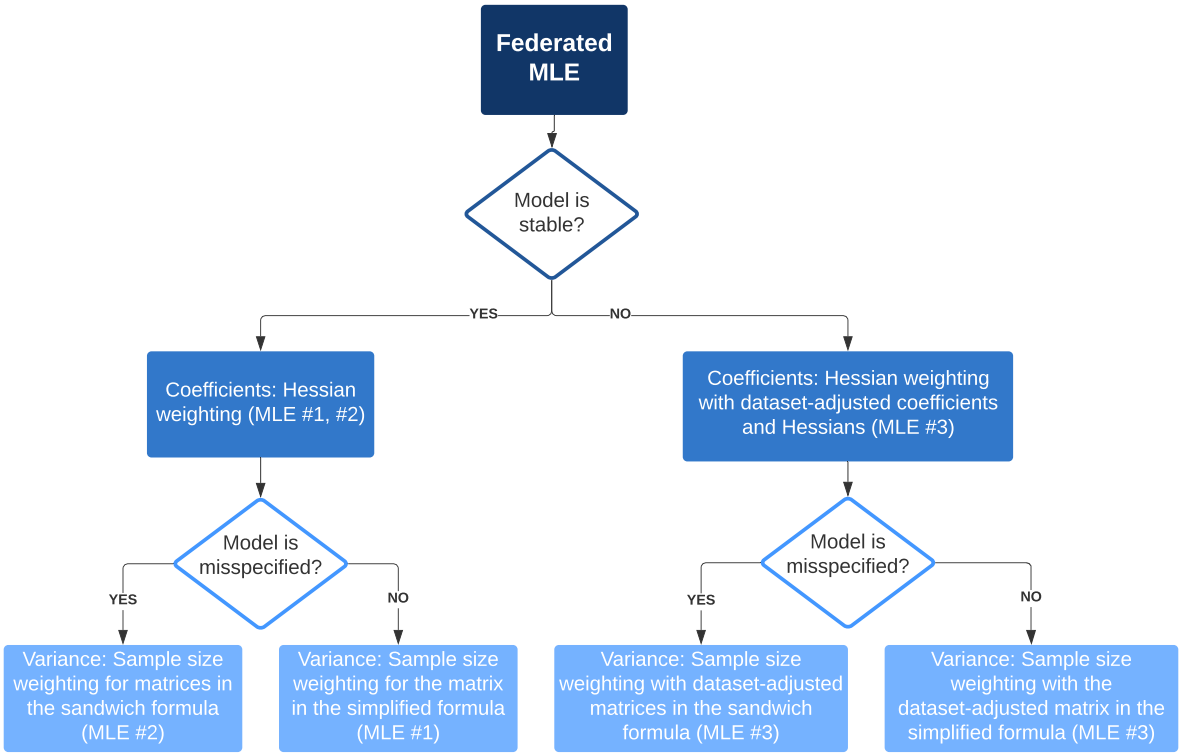
3.1.2 Unrestricted Estimator for Unstable Models (Either Condition 5 or 6 is violated)

In this subsection, we use the outcome model as an example to illustrate our federated maximum likelihood estimator under unstable outcome models (i.e., Condition 6 is violated), but the estimator works similarly for the unstable propensity model (i.e., Condition 5 is violated).

When the outcome model is unstable, but there are some shared parameters across data sets, federating outcome models can still increase the precision of the shared parameters in the estimation. Therefore, for each data set k, we separate the shared parameters from the dataset-specific parameters. That is, $\boldsymbol{\beta}^{(k)} = \left(\boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{\beta}^{(k)}_{\mathcal{S}^{\complement}_k}\right)$, where $\mathcal{S}$ is the set of indices in $\boldsymbol{\beta}^{(k)}$ that are shared (i.e., the corresponding entries in $\boldsymbol{\beta}^{(k)}$ have the same value) across all data sets and $\mathcal{S}^{\complement}_k$ is the set of indices with dataset-specific parameters.[6] For example, in medical applications such as Koenecke et al. (2021), $\boldsymbol{\beta}_{\mathcal{S}}$ could incorporate the coefficient of the treatment that we want to estimate as precisely as possible, while nuisance variables such as dummy variables for non-overlapping fiscal

[6] As a special case, if $\mathcal{S}^{\complement}_k = \emptyset$ for all k, then the outcome model is stable across data sets, but in many applications, $\mathcal{S}^{\complement}_k \neq \emptyset$ for some k.

Figure 2: Flowchart for Federated Maximum Likelihood Estimator and Variance



This flowchart shows the procedure to federate coefficients and variance for either the propensity or outcome model based on MLE. The left and right branches of the "Model is stable?" node correspond to the restricted and unrestricted federated MLE methods, respectively. More details are provided in Sections 3.1.1 and 3.1.2, and in Table 2 (where MLE #1-3 correspond to the three columns in Table 2). Theoretical guarantees are provided in Section 4.1.

years in different data sets could be in $\boldsymbol{\beta}^{(k)}_{\mathcal{S}^{\complement}_k}$.[7] Our federated point estimator aims to combine the estimated $\boldsymbol{\beta}_{\mathcal{S}}$ across all data sets so that we can increase the statistical power to estimate $\boldsymbol{\beta}_{\mathcal{S}}$, while leaving the data set-specific parameters $\boldsymbol{\beta}^{(k)}_{\mathcal{S}^{\complement}_k}$ as they are in the federated estimator.

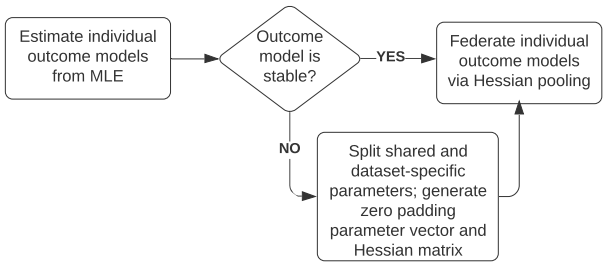
In the outcome model of the combined, individual-level data, we have a larger set of parameters $\boldsymbol{\beta}^{\text{pool}} = (\boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{\beta}^{1}_{\mathcal{S}^{\complement}_1}, \boldsymbol{\beta}^{2}_{\mathcal{S}^{\complement}_2}, \cdots, \boldsymbol{\beta}^{D}_{\mathcal{S}^{\complement}_D})$. For the simplicity of presenting our unrestricted estimator, we assume there are no shared parameters across only a subset of data sets, but our estimator can be easily generalized to the opposite case.[8]

There are three steps to estimating $\boldsymbol{\beta}^{\text{pool}}$ using only the summary level information as shown in Figure 3. The procedure is closely connected to the restricted federated MLE, with the first and third steps being identical. The difference is the addition of the second step compared to the

[7] In more detail, the first data set in Koenecke et al. (2021) includes patient data from fiscal year 2004 through 2016, and the second one includes patient data from fiscal year 2004 through 2019. We include dummy variables for fiscal year, so for the second data set $\mathcal{S}^{\complement}_2 \neq \emptyset$.

[8] If there are some shared parameters across several but not all data sets, we would just need to combine these parameters in $\boldsymbol{\beta}^{\text{pool}}$. For example, if $\boldsymbol{\beta}^{(j)}_{\mathcal{S}^{\complement}}$ and $\boldsymbol{\beta}^{(k)}_{\mathcal{S}^{\complement}}$ are the same for some data sets j and k, then we merge $\boldsymbol{\beta}^{(j)}_{\mathcal{S}^{\complement}}$ and $\boldsymbol{\beta}^{(k)}_{\mathcal{S}^{\complement}}$ in $\boldsymbol{\beta}^{\text{pool}}$.

Figure 3: Flowchart for Federated MLE with Potentially Unstable Outcome Model



This flowchart summarizes the restricted federated MLE in Section 3.1.1 ("YES" at the decision node implies Condition 6 is satisfied), and the unrestricted federated MLE in Section 3.1.2 ("NO" at the decision node implies Condition 6 is violated).

restricted federated estimator. First, we estimate $\boldsymbol{\beta}^{(k)} = \left(\boldsymbol{\beta}_{\mathcal{S}}, \boldsymbol{\beta}^{(k)}_{\mathcal{S}_k^{\complement}}\right)$ by MLE for each individual data set k. Second, we pad the estimated $\hat{\boldsymbol{\beta}}^{(k)}$ with zeros to construct a larger vector, denoted as $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$, such that (a) $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$ has the same dimension as $\boldsymbol{\beta}^{\text{pool}}$ and (b) the parameters that are relevant to data set k are aligned in $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$ and $\boldsymbol{\beta}^{\text{pool}}$. We pad the Hessian $\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}}$ with zeros in the same manner. In the case with no shared parameters across only a subset of data sets, $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$ and $\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}}$ take the form of

$$\hat{\boldsymbol{\beta}}^{\text{pad},(k)} = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{S}} \\ \mathbf{0}_{S_1^{k-1}\times 1} \\ \hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{S}^{\complement}} \\ \mathbf{0}_{S_{k+1}^{D}\times 1} \end{pmatrix}, \quad \hat{\mathbf{H}}^{\text{pad},(k)}_{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}},\boldsymbol{\beta}_{\mathcal{S}}} & \mathbf{0}_{|\mathcal{S}|\times S_1^{k-1}} & \hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}},\boldsymbol{\beta}_{\mathcal{S}^{\complement}}} & \mathbf{0}_{|\mathcal{S}|\times S_{k+1}^{D}} \\ \mathbf{0}_{S_1^{k-1}\times|\mathcal{S}|} & \mathbf{0}_{S_1^{k-1}\times S_1^{k-1}} & \mathbf{0}_{S_1^{k-1}\times S_k^{k}} & \mathbf{0}_{S_1^{k-1}\times S_{k+1}^{D}} \\ \hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}^{\complement}},\boldsymbol{\beta}_{\mathcal{S}}} & \mathbf{0}_{S_k^{k}\times S_1^{k-1}} & \hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}^{\complement}},\boldsymbol{\beta}_{\mathcal{S}^{\complement}}} & \mathbf{0}_{S_k^{k}\times S_{k+1}^{D}} \\ \mathbf{0}_{S_{k+1}^{D}\times|\mathcal{S}|} & \mathbf{0}_{S_{k+1}^{D}\times S_1^{k-1}} & \mathbf{0}_{S_{k+1}^{D}\times S_k^{k}} & \mathbf{0}_{S_{k+1}^{D}\times S_{k+1}^{D}} \end{pmatrix}, \tag{13}$$

where $S_{j_1}^{j_2} = \sum_{j=j_1}^{j_2} |\mathcal{S}_j^{\complement}|$, $\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}},\boldsymbol{\beta}_{\mathcal{S}}} = \frac{\partial^2 \boldsymbol{\ell}_{n_k}(\hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{S}})}{\partial \boldsymbol{\beta}^{(k)}_{\mathcal{S}} (\partial \boldsymbol{\beta}^{(k)}_{\mathcal{S}})^{\top}}$, $\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}},\boldsymbol{\beta}_{\mathcal{S}^{\complement}}} = \frac{\partial^2 \boldsymbol{\ell}_{n_k}(\hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{S}})}{\partial \boldsymbol{\beta}^{(k)}_{\mathcal{S}} (\partial \boldsymbol{\beta}^{(k)}_{\mathcal{S}^{\complement}})^{\top}} = \left(\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}^{\complement}},\boldsymbol{\beta}_{\mathcal{S}}}\right)^{\top}$, $\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}^{\complement}},\boldsymbol{\beta}_{\mathcal{S}^{\complement}}} = \frac{\partial^2 \boldsymbol{\ell}_{n_k}(\hat{\boldsymbol{\beta}}^{(k)}_{\mathcal{S}^{\complement}})}{\partial \boldsymbol{\beta}^{(k)}_{\mathcal{S}} (\partial \boldsymbol{\beta}^{(k)}_{\mathcal{S}^{\complement}})^{\top}}$, and $\mathbf{0}_{n_1\times n_2}$ represents an $n_1 \times n_2$ matrix of zeros for any positive integers n_1 and n_2. The last step is to use Hessian weighting (9), where the Hessian is $\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}_{\mathcal{S}},\boldsymbol{\beta}_{\mathcal{S}}}$.

There are also three steps to estimating the unrestricted federated variance. The procedure connects to the restricted federated variance estimator in a fashion similar to the federated point estimator. We first estimate $\mathbf{A}^{(k)}_{\boldsymbol{\beta}}$ (and $\mathbf{B}^{(k)}_{\boldsymbol{\beta}}$ if the outcome model is misspecified, i.e., Condition 8 is violated). Next, we pad $\mathbf{A}^{(k)}_{\boldsymbol{\beta}}$ (and $\mathbf{B}^{(k)}_{\boldsymbol{\beta}}$ if necessary) in the same manner as $\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta}}$ to obtain $\hat{\mathbf{A}}^{\text{pad},(k)}_{\boldsymbol{\beta}}$ (and $\hat{\mathbf{B}}^{\text{pad},(k)}_{\boldsymbol{\beta}}$ if necessary), as shown in Eq. (13). Finally, we use sample size weighting (10) to combine $\hat{\mathbf{A}}^{\text{pad},(k)}_{\boldsymbol{\beta}}$ (and $\hat{\mathbf{B}}^{\text{pad},(k)}_{\boldsymbol{\beta}}$ if necessary).

As a caveat, for the stable outcome model (Condition 6), we could artificially treat some shared parameters as dataset-specific parameters and use the unrestricted federated estimator proposed in this subsection. However, we do not suggest doing so because a flexible pooled outcome model

Table 2: Federated Maximum Likelihood Estimator

Description	Assume Stable Outcome Model (MLE #1)	Assume Stable Misspecified Outcome Model (MLE #2)	Assume Unstable Outcome Model (MLE #3)
Stable Covariate Distribution	yes or no	yes or no	yes or no
Stable Outcome Model	yes	yes	no
Correct Outcome Model Specification	yes	no	yes or no
Sample Size Assumption	yes or no	yes or no	yes
Coefficient $\boldsymbol{\beta}$ federation	Hessian weighting: $\left(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\right)^{-1}\left(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\hat{\boldsymbol{\beta}}^{(k)}\right)$		Hessian weighting with $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$ and $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}$
Variance $\mathbf{V}_{\boldsymbol{\beta}}$ federation	$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta}}^{-1}$ Federate the estimated $\mathbf{A}_{\boldsymbol{\beta}}$ (and $\mathbf{B}_{\boldsymbol{\beta}}$ if necessary) by sample size weighting.	$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta}}^{-1}\mathbf{B}_{\boldsymbol{\beta}}\mathbf{A}_{\boldsymbol{\beta}}^{-1}$	Federate the estimated $\mathbf{A}_{\boldsymbol{\beta}}^{\text{pad}}$ (and $\mathbf{B}_{\boldsymbol{\beta}}^{\text{pad}}$ if necessary) by sample size weighting.
Asymptotic Results	Theorem 1	Proposition 2	Proposition 3

This table states the federation procedure for coefficients and variance in the outcome model from MLE under various conditions. The second to fifth rows correspond to Conditions 3, 6 and 8, and Assumption 2, respectively. We use "yes or no" to indicate the case where the solution does not vary with whether or not the condition is satisfied. Note that this table holds when we use MLE to estimate the propensity model as well. In this table, $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} = \frac{\partial^2 \boldsymbol{\ell}_{n_k}(\hat{\boldsymbol{\beta}}^{(k)})}{\partial \boldsymbol{\beta}^{(k)} (\partial \boldsymbol{\beta}^{(k)})^\top}$ denotes the Hessian, and the definition of $\mathbf{A}_{\boldsymbol{\beta}}$ and $\mathbf{B}_{\boldsymbol{\beta}}$ can be found in Table 1. $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}$ scales with sample size n_k, while $\mathbf{A}_{\boldsymbol{\beta}}$ and $\mathbf{B}_{\boldsymbol{\beta}}$ do not. When the outcome model is stable across data sets (Condition 6 holds), the coefficient federation formula is the same for all scenarios, and the variance federation formula depends on whether the information matrix equivalence (Condition 8) holds. When the outcome model is unstable (Condition 6 is violated), we use the same federation scheme as the stable outcome model, but with $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}$, $\mathbf{A}_{\boldsymbol{\beta}}^{\text{pad}}$, and $\mathbf{B}_{\boldsymbol{\beta}}^{\text{pad}}$ that pad $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}$, $\mathbf{A}_{\boldsymbol{\beta}}$, and $\mathbf{B}_{\boldsymbol{\beta}}$ with zeros as shown in Section 3.1.2.

with more parameters may lead to a less efficient federated estimator compared to that from the correctly specified (most parsimonious) model. We provide a formal statement for generalized linear models in the following proposition.

Proposition 1. *Suppose Y_i follows a generalized linear model, and the true parameters $\boldsymbol{\beta}_0^{(k)}$ are the same for all k. If $\mathcal{S}_k^{\complement} \neq \emptyset$ and we use the unrestricted federated estimator in this subsection to estimate $\boldsymbol{\beta}^{\text{pool}}$, then we get a weakly less efficient estimate of $\boldsymbol{\beta}_{\mathcal{S}}$ than that from the restricted federated estimator in Section 3.1.1.*

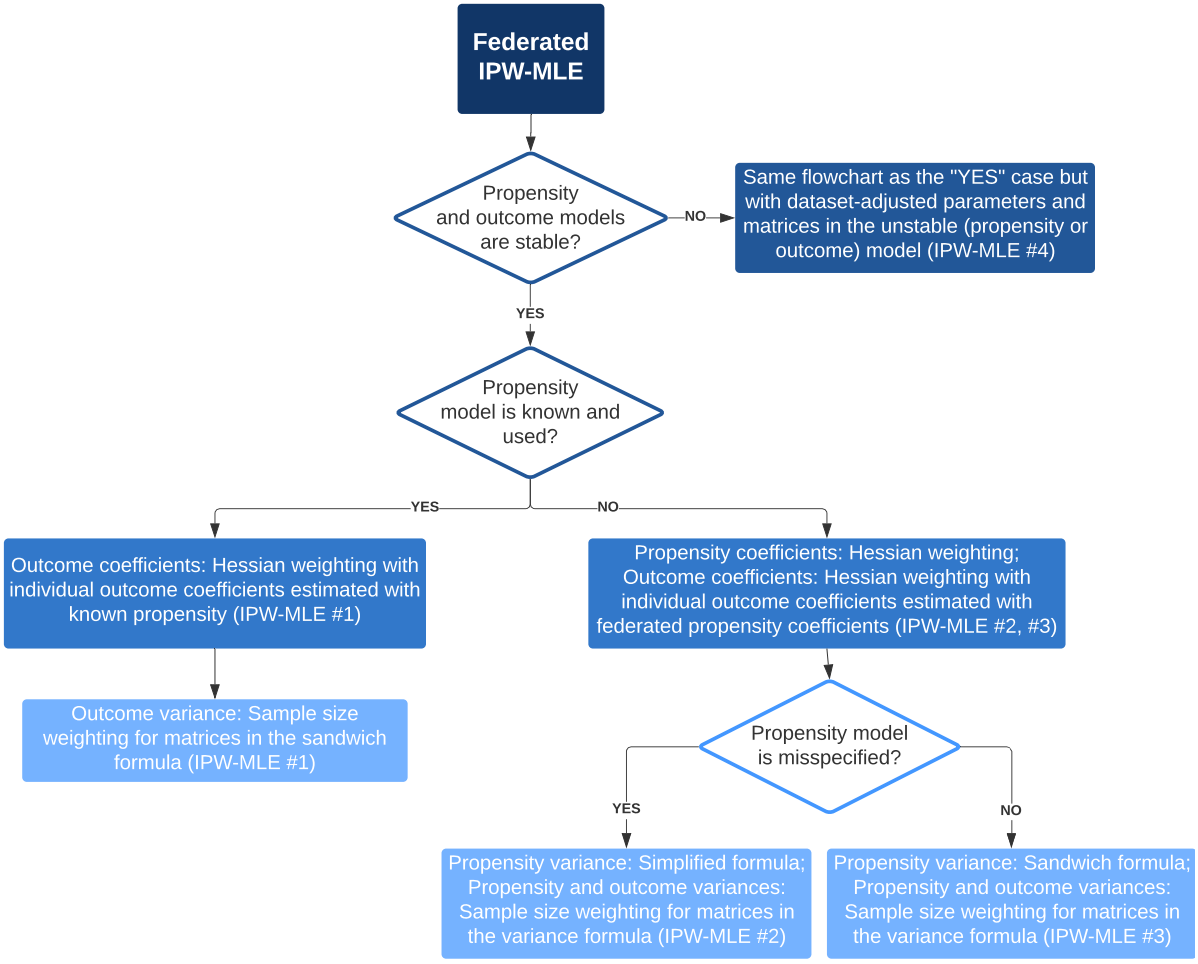
3.2 Federated Inverse Propensity-Weighted Maximum Likelihood Estimator

3.2.1 Restricted Estimator for Stable Models (Conditions 5 and 6 hold)

IPW-MLE uses propensity scores in the estimation of parameters in the outcome model, so compared with federated MLE, we need to additionally consider the federation of parameters in the propensity models in the federated IPW-MLE. We focus on the setting in which both the propensity and outcome models on individual data sets are estimated from MLE. Then we can leverage our advances in the federated MLE estimator (in Section 3.1.1) and its asymptotic properties (in

Section 4.1) to provide a theoretically guaranteed federated IPW-MLE.

Figure 4: Flowchart for Federated Inverse Propensity-Weighted Maximum Likelihood Estimator and Variance



This flowchart shows the procedure to federate coefficients and variance in the outcome model and, if estimated, the propensity model based on IPW-MLE. The left and right branches of the "Propensity and outcome models are stable?" node correspond to restricted and unrestricted federated IPW-MLE methods, respectively. More details are provided in Sections 3.2.1 and 3.2.2, and in Table 3 (where IPW-MLE #1-4 correspond to the four columns in Table 3). Theoretical guarantees are provided in Section 4.2.

Let us first consider the federated point estimator. In the simple case where the true propensity score is known and used (Condition 4), we do not need to pool the propensity scores across data sets. On the other hand, for a more complicated case where the propensity score is estimated from a parametric model using MLE, we use the federated MLE point estimator to estimate the coefficients of the pooled propensity model. Specifically, we use Hessian weighting (9) to combine the estimated coefficients of the propensity model on individual data sets.

After we estimate the federated coefficients of the propensity model, the next step is to estimate the federated outcome model. To estimate the federated outcome model, we first use the federated propensity scores to estimate the coefficients of the outcome model for each individual data set.

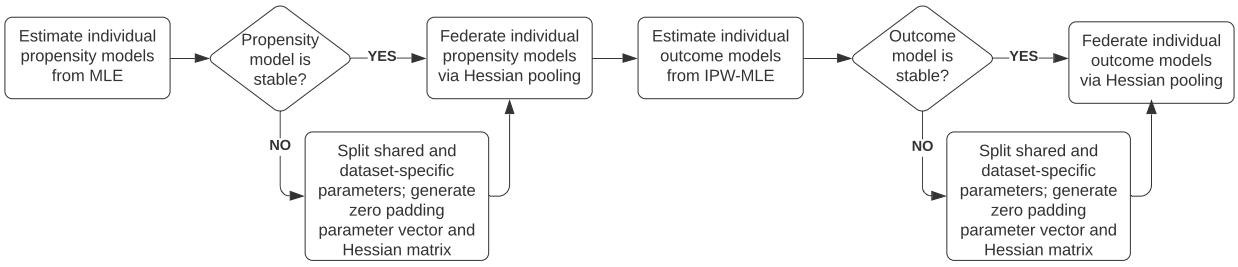
We then use Hessian weighting (9) to combine together coefficients of the outcome model across all data sets.

The federated variance estimator is based on, and analogous to, the asymptotic variance $\mathbf{V}_{\boldsymbol{\beta}}$ of the IPW-MLE on a single data set. When the true propensity score is known and used (Condition 4), $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}\mathbf{D}_{\boldsymbol{\beta},\varpi}\mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}$. When the propensity score is estimated, $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}(\mathbf{D}_{\boldsymbol{\beta},\varpi} - \mathbf{C}_{\boldsymbol{\beta},\varpi}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}_{\boldsymbol{\beta},\varpi}^{\top})\mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}$ for ATE weighting, and $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}(\mathbf{D}_{\boldsymbol{\beta},\varpi} - \mathbf{C}_{\boldsymbol{\beta},\varpi,1}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}_{\boldsymbol{\beta},\varpi,2}^{\top} - \mathbf{C}_{\boldsymbol{\beta},\varpi,2}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}_{\boldsymbol{\beta},\varpi,1}^{\top} + \mathbf{C}_{\boldsymbol{\beta},\varpi,2}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}_{\boldsymbol{\beta},\varpi,2}^{\top})\mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}$ for ATT weighting, where $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}_{\boldsymbol{\gamma}}^{-1}$ if the propensity model is correctly specified (Condition 4 is violated and Condition 7 holds), and $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}_{\boldsymbol{\gamma}}^{-1}\mathbf{B}_{\boldsymbol{\gamma}}\mathbf{A}_{\boldsymbol{\gamma}}^{-1}$ otherwise (Conditions 4 and 7 are violated).[9] The definitions of $\mathbf{A}_{\boldsymbol{\beta},\varpi}$, $\mathbf{D}_{\boldsymbol{\beta},\varpi}$, $\mathbf{C}_{\boldsymbol{\beta},\varpi}$, $\mathbf{C}_{\boldsymbol{\beta},\varpi,1}$, $\mathbf{C}_{\boldsymbol{\beta},\varpi,2}$, $\mathbf{A}_{\boldsymbol{\gamma}}$ and $\mathbf{B}_{\boldsymbol{\gamma}}$ can be found in Table 1. To estimate the federated variance, we first estimate $\mathbf{A}_{\boldsymbol{\beta},\varpi}$ and $\mathbf{D}_{\boldsymbol{\beta},\varpi}$ (as well as $\mathbf{C}_{\boldsymbol{\beta},\varpi}$, $\mathbf{A}_{\boldsymbol{\gamma}}$, and $\mathbf{B}_{\boldsymbol{\gamma}}$ if necessary) on each individual data set, and then use sample size weighting (10) to combine $\mathbf{A}_{\boldsymbol{\beta},\varpi}$ and $\mathbf{D}_{\boldsymbol{\beta},\varpi}$ (as well as $\mathbf{C}_{\boldsymbol{\beta},\varpi}$, $\mathbf{A}_{\boldsymbol{\gamma}}$, and $\mathbf{B}_{\boldsymbol{\gamma}}$ if necessary) across all data sets.

3.2.2 Unrestricted Estimator for Unstable Models (either Condition 5 or 6 is violated)

Our unrestricted federated IPW-MLE is built on our restricted federated estimator in Section 3.2.1. Most steps are the same, with the addition of one or two steps to deal with the dataset-specific parameters in the propensity and/or outcome models as shown in Figure 5. As with the restricted federated IPW-MLE, we focus on the setting in which the propensity model is estimated from MLE.

Figure 5: Flowchart for Federated IPW-MLE with Estimated Propensity Scores and Potentially Unstable Propensity and/or Outcome Models



This flowchart summarizes the federated IPW-MLE for stable propensity and outcome models (where Conditions 5 and 6 are satisfied, i.e. "YES" for both decision nodes) in Section 3.2.1, and the federated IPW-MLE for either unstable propensity or unstable outcome model in Section 3.2.2 (where Conditions 5 and/or 6 are violated, i.e. "NO" for one or both decision nodes). This flowchart focuses on the setting in which the propensity scores are estimated (using MLE). If the true propensity scores are known and used, we skip the federation of individual propensity models.

[9]The expression of $\mathbf{V}_{\boldsymbol{\beta}}$ for the ATE weights is provided in Wooldridge (2002, 2007) when the true propensity score is used or when the estimated propensity model is correctly specified. We extend Wooldridge (2002, 2007) and provide the expression for $\mathbf{V}_{\boldsymbol{\beta}}$ for the misspecified propensity model and for the ATT weights in Lemma 1 in Section 4.

Table 3: Federated Inverse Propensity-Weighted Maximum Likelihood Estimator

Description	Assume Stable Known Propensity and Stable Outcome Model (IPW-MLE #1)	Assume Stable Propensity and Stable Outcome Model (IPW-MLE #2)	Assume Stable Misspecified Propensity and Stable Outcome Model (IPW-MLE #3)	Assume Unstable Propensity or Unstable Outcome Model (IPW-MLE #4)
Stable Covariate Distribution	yes or no	yes or no	yes or no	yes or no
Known Propensity	yes	no	no	yes or no
Stable Propensity Model	yes	yes	yes	yes or no
Stable Outcome Model	yes	yes	yes	yes or no
Correct Propensity Model Specification	yes	yes	no	yes or no
Correct Outcome Model Specification	yes or no	yes or no	yes or no	yes or no
Sample Size Assumption	yes or no	yes or no	yes or no	yes
Coefficient $\boldsymbol{\beta}$ federation	(1) Estimate $\hat{\boldsymbol{\beta}}^{(k)}$ using $\boldsymbol{\gamma}_0$; (2) Federate $\hat{\boldsymbol{\beta}}^{(k)}$ by Hessian weighting.	(1) Federate $\hat{\boldsymbol{\gamma}}^{(k)}$ by Hessian weighting; (2) Estimate $\hat{\boldsymbol{\beta}}^{(k)}$ using $\hat{\boldsymbol{\gamma}}^{\text{fed}}$; (3) Federate $\hat{\boldsymbol{\beta}}^{(k)}$ by Hessian weighting.		Same federation procedure with $\hat{\boldsymbol{\gamma}}^{\text{pad},(k)}$ and $\hat{\mathbf{H}}_{\boldsymbol{\gamma}}^{\text{pad},(k)}$ if propensity models are unstable and estimated, and with $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$ and $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}$ if outcomes models are unstable
Variance $\mathbf{V}_{\boldsymbol{\beta}}$ federation	$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}\mathbf{D}_{\boldsymbol{\beta},\varpi}\mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}$	$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}(\mathbf{D}_{\boldsymbol{\beta},\varpi} - \mathbf{M}_{\boldsymbol{\beta},\varpi,\boldsymbol{\gamma}})\mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}$ $\mathbf{M}_{\boldsymbol{\beta},\varpi,\boldsymbol{\gamma}} = \mathbf{C}_{\boldsymbol{\beta},\varpi}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}_{\boldsymbol{\beta},\varpi}^{\top}$ for ATE weighting; $\mathbf{M}_{\boldsymbol{\beta},\varpi,\boldsymbol{\gamma}} = \mathbf{C}_{\boldsymbol{\beta},\varpi,1}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}_{\boldsymbol{\beta},\varpi,2}^{\top} + \mathbf{C}_{\boldsymbol{\beta},\varpi,2}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}_{\boldsymbol{\beta},\varpi,1}^{\top} - \mathbf{C}_{\boldsymbol{\beta},\varpi,2}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}_{\boldsymbol{\beta},\varpi,2}^{\top}$ for ATT weighting		Same federation procedure with $\hat{\boldsymbol{\gamma}}^{\text{pad},(k)}$, estimated $\mathbf{A}_{\boldsymbol{\gamma}}^{\text{pad},(k)}$, $\mathbf{C}_{\boldsymbol{\beta},\varpi}^{\text{pad},(k)}$ (and $\mathbf{B}_{\boldsymbol{\gamma}}^{\text{pad},(k)}$ if needed) if propensity models are unstable and estimated, and with $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$, estimated $\mathbf{A}_{\boldsymbol{\beta},\varpi}^{\text{pad},(k)}$, $\mathbf{D}_{\boldsymbol{\beta},\varpi}^{\text{pad},(k)}$, $\mathbf{C}_{\boldsymbol{\beta},\varpi}^{\text{pad},(k)}$ if outcomes models are unstable
		$\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}_{\boldsymbol{\gamma}}^{-1}$	$\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}_{\boldsymbol{\gamma}}^{-1}\mathbf{B}_{\boldsymbol{\gamma}}\mathbf{A}_{\boldsymbol{\gamma}}^{-1}$	
	(1) Federate $\hat{\boldsymbol{\gamma}}^{(k)}$ by Hessian weighting (skip for known propensity); (2) Estimate $\boldsymbol{\beta}^{(k)}$ using $\hat{\boldsymbol{\gamma}}^{\text{fed}}$; (3) Estimate $\mathbf{A}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{C}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{D}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{A}_{\boldsymbol{\gamma}}^{(k)}$, and $\mathbf{B}_{\boldsymbol{\gamma}}^{(k)}$ using $\hat{\boldsymbol{\gamma}}^{\text{fed}}$ and $\hat{\boldsymbol{\beta}}^{\text{fed}}$; (4) Federate estimated $\mathbf{A}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{C}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{D}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{A}_{\boldsymbol{\gamma}}^{(k)}$, and $\mathbf{B}_{\boldsymbol{\gamma}}^{(k)}$ by sample size weighting.		
Results	Theorem 2			Proposition 4

This table states the federation procedure for coefficients and variance in the outcome model from IPW-MLE under various conditions. The second to eighth rows correspond to Conditions 3-8 and Assumption 2, respectively. The definitions of $\mathbf{A}_{\boldsymbol{\beta},\varpi}, \mathbf{D}_{\boldsymbol{\beta},\varpi}, \mathbf{C}_{\boldsymbol{\beta},\varpi}, \mathbf{C}_{\boldsymbol{\beta},\varpi,1}, \mathbf{C}_{\boldsymbol{\beta},\varpi,2}, \mathbf{A}_{\boldsymbol{\gamma}}$, and $\mathbf{B}_{\boldsymbol{\gamma}}$ can be found in Table 1. When the propensity model is estimated (Condition 4 is violated), the coefficient federation procedure is the same for all scenarios, but is simplified when the true propensity is used (Condition 4 holds). We provide the variance federation formula for the ATE and ATT weights that depends on whether we use the true propensity (Condition 4) and whether the estimated propensity model is correctly specified (Condition 7). The definitions of ATE and ATT weights can be found in Section 2.2.2. $\hat{\boldsymbol{\gamma}}^{\text{pad},(k)}$, $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$, $\mathbf{A}_{\boldsymbol{\beta},\varpi}^{\text{pad},(k)}$, $\mathbf{D}_{\boldsymbol{\beta},\varpi}^{\text{pad},(k)}$,$\mathbf{C}_{\boldsymbol{\beta},\varpi}^{\text{pad},(k)}$,$\mathbf{C}_{\boldsymbol{\beta},\varpi,1}^{\text{pad},(k)}$,$\mathbf{C}_{\boldsymbol{\beta},\varpi,2}^{\text{pad},(k)}$, $\mathbf{A}_{\boldsymbol{\gamma}}^{\text{pad},(k)}$, and $\mathbf{B}_{\boldsymbol{\gamma}}^{\text{pad},(k)}$ are $\hat{\boldsymbol{\gamma}}^{(k)}$, $\hat{\boldsymbol{\beta}}^{(k)}$ $\mathbf{A}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{D}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{C}_{\boldsymbol{\beta},\varpi}^{(k)}, \mathbf{C}_{\boldsymbol{\beta},\varpi,1}^{(k)}, \mathbf{C}_{\boldsymbol{\beta},\varpi,2}^{(k)}, \mathbf{A}_{\boldsymbol{\gamma}}^{(k)}$, and $\mathbf{B}_{\boldsymbol{\gamma}}^{(k)}$ padded with zeros.

For the federated point estimator, we first need to federate the propensity scores across data sets if they are unknown. Similar to the stable propensity model, we first estimate the individual propensity model for each data set. When the propensity models are unstable (Condition 5 is violated), we use a similar approach as in Section 3.1.2 to federate propensity models. Specifically, we first separate the stable parameters from the unstable parameters for each data set. Second, we specify a larger set of parameters for the pooled propensity model and pad the estimated parameters and Hessian from each data set with zeros to be aligned with the dimension of those for the pooled model. Third, we use Hessian weighting to combine the zero-padded, estimated parameters across data sets together.

Next, we estimate the federated outcome model using the individual outcome models estimated from the federated propensity model. If the outcome models are unstable (i.e., Condition 6 is violated), we follow the same three steps as the unstable propensity models to combine parameters in individual outcome models.

For the unrestricted federated variance estimator, we use the same procedure as the restricted federated variance estimator in Section 3.2.1 and Figure 4, but we pad each matrix used in the federation procedure with zeros in the same fashion as we pad the Hessian matrix.

3.3 Federated AIPW Estimator

3.3.1 Restricted Estimator for Stable Models (Conditions 5 and 6 hold)

The AIPW estimator uses both outcome and propensity models, so we need to consider the federation of outcome models as well as the federation of propensity models if the propensity score is estimated.

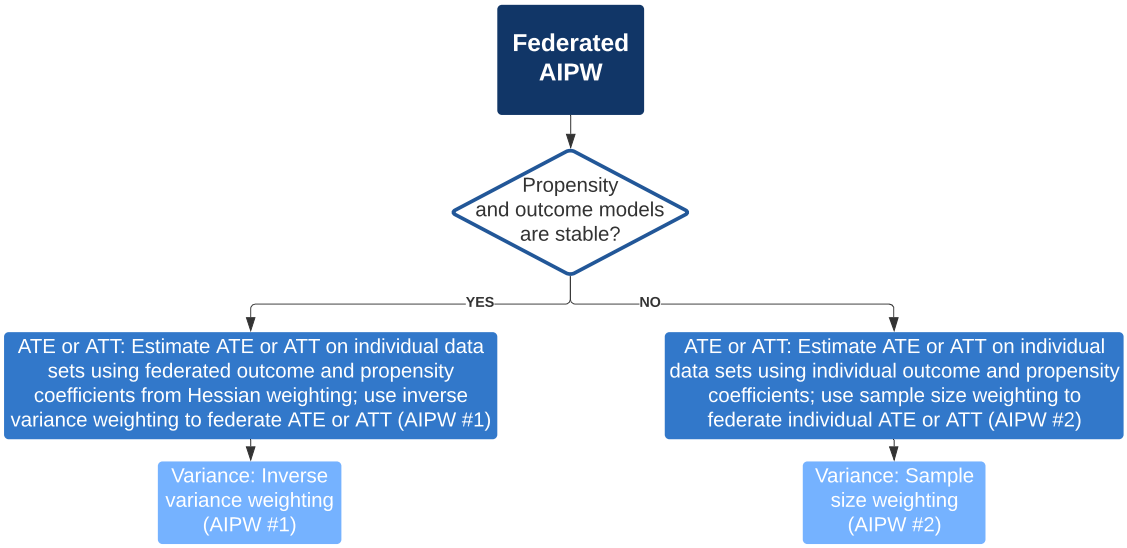
There are three steps to estimating the federated ATE or federated ATT. We first federate the outcome models (and propensity models if necessary) across all data sets. If propensity and/or outcome models are estimated from the MLE, then we use Hessian weighting (9), the same as how we combine outcome models (and propensity models if necessary) in the federated MLE estimator.

Next, we use the federated/true propensity, and federated outcome models to estimate the ATE or ATT by the AIPW estimator on individual data sets. Finally, we use inverse variance weighting (11) to federate ATE or ATT from the AIPW estimator across all data sets. For the variance of the federated ATE or ATT, we use inverse variance weighting (11) to federate the variance of ATE or ATT across all data sets.

It is worth noting that these procedures to federate ATE or ATT and its variance are quite general. They are robust to propensity or outcome model misspecification (i.e., where either Condition 7 or 8 is violated).[10] If propensity and/or outcome models are estimated from other approaches (e.g., random forests (Wager and Athey, 2018)), we only need to modify the approach to combine individual propensity and outcome models in the first step of the federation procedure of ATE or

[10] If both Conditions 7 and 8 are violated, we cannot consistently estimate ATE or ATT, even in the classical setting where we only have one data set; therefore, this scenario is not our main focus.

Figure 6: Flowchart for Federated AIPW Estimator and Variance



This flowchart shows the procedure to federate ATE or ATT, and variance, based on AIPW. The left and right branches of the "Propensity and outcome models are stable?" node correspond to the restricted and unrestricted federated AIPW methods, respectively. More details are provided in Sections 3.3.1 and 3.3.2, and in Table 4 (where AIPW #1-2 correspond to the two columns in Table 4). Theoretical guarantees are provided in Section 4.3.

ATT. However, federating other approaches and providing the theoretical guarantees are beyond the scope of this paper and are left for future work.

3.3.2 Unrestricted Estimator for Unstable Models (either Condition 5 or 6 is violated)

If either the parameters in the propensity model or parameters in the outcome model vary with the data set, then the score $\hat{\phi}(\mathbf{X}_i, W_i, Y_i)$, which is a function of propensity and outcome models in the definition of the AIPW estimator, also varies with the data set. We therefore propose to estimate the federated ATE or ATT using sample size weighting to combine the individual ATE or ATT. That is,

$$\hat{\tau}^{\text{fed}}_{\text{aipw}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \hat{\tau}^{(k)}_{\text{aipw}} \qquad \hat{\mathbf{V}}^{\text{fed}}_{\tau} = \sum_{k=1}^{D} \frac{n_k^2}{n_{\text{pool}}} \hat{\mathbf{V}}^{(k)}_{\tau}, \tag{14}$$

where $\hat{\tau}^{(k)}_{\text{aipw}}$ is estimated from individual propensity and outcome models on data set k.

Sample size weighting (14) is quite flexible and allows parameters in the propensity or outcome model to vary arbitrarily across data sets. The tradeoff is that, in the case of stable propensity and outcome models (Conditions 5 and 6), sample size weighting is less efficient than the inverse variance weighting used in Section 3.3.1, because inverse variance weighting has the smallest variance among all weighted averages.

Table 4: Federated AIPW Estimator

Description	Assume Stable Propensity and Stable Outcome Model (AIPW #1)	Assume Unstable Propensity or Unstable Outcome Model (AIPW #2)
Stable Covariate Distribution	yes or no	yes or no
True Propensity	yes or no	yes or no
Stable Propensity Model	yes	yes or no
Stable Outcome Model	yes	yes or no
Correct Propensity Model Specification	yes or no	yes or no
Correct Outcome Model Specification	yes or no	yes or no
Sample Size Assumption	no	yes
ATE or ATT τ federation	(1) Federate $\hat{\boldsymbol{\beta}}^{(k)}$ (and $\hat{\boldsymbol{\gamma}}^{(k)}$) by Hessian weighting; (2) Estimate $\tau^{(k)}$ using $\hat{\boldsymbol{\beta}}^{\text{fed}}$ and $\hat{\boldsymbol{\gamma}}^{\text{fed}}$ (or $\boldsymbol{\gamma}_0^{(k)}$ if known); (3) Federate $\hat{\tau}^{(k)}$ by inverse variance weighting.	(1) Estimate $\tau^{(k)}$ using $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{\gamma}}^{(k)}$ (or $\boldsymbol{\gamma}_0^{(k)}$ if known); (2) Federate $\hat{\tau}^{(k)}$ by sample size weighting.
Variance $\mathbf{V}_\tau$ federation	Inverse variance weighting	Sample size weighting
Results	Theorem 3	

This table states the federation procedure of the AIPW estimator for ATE or ATT and its variance under various conditions. The second to eighth rows correspond to Conditions 3-8 and Assumption 2. Under parametric propensity and outcome models (Conditions 1 and 2), these two models can potentially be misspecified. As long as one is correctly specified (either Condition 7 or 8 holds), τ can be consistently estimated. When the propensity and outcome models are the same across all data sets (under Conditions 5 and 6), we federate the propensity and outcome models before estimating ATE or ATT on individual data sets. Otherwise (when either Condition 5 or 6 is violated), we leave them as they are when estimating the individual ATE or ATT. For the stable propensity and outcome models, we federate individual ATE or ATT by inverse variance weighting, which is generally more efficient than the sample size weighting used for the unstable propensity or outcome model (when either Condition 5 or 6 is violated).

4 Asymptotic Results

In this section, we derive the asymptotic distributions of our federated MLE, IPW-MLE, and AIPW estimators. The asymptotic distributions correspond to their respective estimators using the combined, individual-level data. In addition, we show that our federated variance estimators are consistent, which allows us to construct valid confidence intervals for treatment effect estimates and other parameters of interest. Our asymptotic results are confirmed, and their finite-sample properties are demonstrated through simulations in Appendix D. All proofs are provided in Appendix E. In the following notation, the superscript "fed" represents the federated estimator and the superscript "pool" represents the estimator from the combined, individual-level data.

4.1 Maximum Likelihood Estimation

We start with showing that when the outcome model is stable across data sets (Condition 6 holds) and correctly specified (Condition 8 holds), the federated coefficients $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}}$ have the same asymptotic

distribution as $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{mle}}$ estimated from the combined, individual-level data. Moreover, both $\hat{\mathbf{V}}^{\text{fed}}_{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}^{\text{pool}}_{\boldsymbol{\beta}}$ are feasible estimators for the asymptotic variance of $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}}$ and $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{mle}}$.

Theorem 1 (Restricted Federated MLE for a Stable and Correctly Specified Outcome Model)**.** *Suppose Assumption 1.1 and Conditions 6 and 8 hold. Furthermore, suppose either one of the following cases hold: (a) $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ is the same across all data sets or (b) Assumption 2 holds and there exists $M < \infty$ such that $\left\|\mathcal{I}^{\text{pool}}(\boldsymbol{\beta})^{-1}\mathcal{I}^{(k)}(\boldsymbol{\beta})\right\|_2 \leq M$. As $n_1, \cdots, n_D \to \infty$, we have*

$$n^{1/2}_{\text{pool}}(\hat{\mathbf{V}}^{\text{pool}}_{\boldsymbol{\beta}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \tag{15}$$

where d is the dimension of $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_0$ represents the true parameters (see Condition 1). If we replace $\hat{\mathbf{V}}^{\text{pool}}_{\boldsymbol{\beta}}$ by $\hat{\mathbf{V}}^{\text{fed}}_{\boldsymbol{\beta}}$ and/or replace $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}}$ by $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{mle}}$, Eq. (15) *continues to hold.*

The convergence rate for $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}}$ and $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{mle}}$ is $n^{1/2}_{\text{pool}}$, which is the optimal rate. Compared to the $n^{1/2}_k$ convergence rate for $\hat{\boldsymbol{\beta}}^{(k)}_{\text{mle}}$, we improve efficiency via federation. $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}}$ and $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{mle}}$ are asymptotically equivalent if the information matrix $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ is the same for all k (e.g., the homogeneous data sets have stable covariate distribution and stable propensity and outcome models – Conditions 3, 5 and 6 hold) . In this case, we can leave the sample size proportion n_k/n_{pool} unrestricted. In the second case of Theorem 1 where $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ varies (e.g., under heterogeneous data sets with unstable covariate distributions – Condition 3 is violated), $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}}$ and $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{mle}}$ can still be asymptotically equivalent if the limit of n_k/n_{pool} exists. The existence of this limit implies that $\mathcal{I}^{\text{pool}}(\boldsymbol{\beta})^{-1}$ can be properly defined.

Note that Theorem 1 holds for the propensity model when we use the federated MLE and the same federation formulas as the outcome model if we replace $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}}$ and $\hat{\mathbf{V}}^{\text{fed}}_{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\gamma}}^{\text{fed}}_{\text{mle}}$ and $\hat{\mathbf{V}}^{\text{fed}}_{\boldsymbol{\gamma}}$, respectively.

If the model is misspecified, MLE does not converge to the true parameters $\boldsymbol{\beta}_0$; however, MLE does converge in probability to the parameters that minimize the Kullback-Leibler Information Criterion between true and misspecified models (White, 1982).[11] These parameters are denoted as $\boldsymbol{\beta}^*$ for the outcome model. We leverage this property of MLE for the misspecified model and show that, with model misspecification (Condition 8 is violated), Theorem 1 continues to hold with $\boldsymbol{\beta}_0$ replaced by $\boldsymbol{\beta}^*$.

Proposition 2 (Restricted Federated MLE for a Stable and Misspecified Outcome Model)**.** *Suppose the regularity conditions in White (1982) hold, Condition 6 holds, Condition 8 is violated, and the parameters $\boldsymbol{\beta}^*$ are the same across all data sets. Then, Theorem 1 continues to hold with $\boldsymbol{\beta}_0$ replaced by $\boldsymbol{\beta}^*$.*

Similar to Theorem 1, Proposition 2 also holds for the misspecified propensity model (where Condition 7 is violated). Since $\boldsymbol{\beta}^*$ is generally different from $\boldsymbol{\beta}_0$, we cannot get a consistent estimate for the treatment coefficient in the outcome model; however, we can use it as the input for the AIPW estimator, which is robust to the outcome model misspecification. Similarly, for the misspecified

[11] More details are presented in Section E.1 in the Appendix.

propensity model, we can use it as the input for the IPW-MLE and AIPW estimators that are robust to propensity model misspecification.

With an unstable outcome model (where Condition 6 is violated), the federated coefficients and variance from our unrestricted federated MLE in Section 3.1.2 are still asymptotically equivalent to those estimated from the combined, individual-level data. The results in Theorem 1 continue to hold for the modified federation procedure under unstable models (Condition 5 is violated), as shown in Proposition 3.

Proposition 3 (Unrestricted Federated MLE for an Unstable Outcome Model)**.** *Suppose Assumptions 1.1 and 2 hold, Condition 6 is violated, and* $\mathbf{A}^{\text{pad},(k)}$ *is full rank for all* k*. If we use the method in Section 3.1.2 to federate coefficients and variance, then Theorem 1 continues to hold.*

Remark 1. When the outcome model is unstable, if we continue using the same federation formulas as those for stable models in Section 3.1.1, Proposition 5 in Appendix A shows that Theorem 1 continues to hold for some special cases, but with $\boldsymbol{\beta}_0$ replaced by some weighted average of $\boldsymbol{\beta}_0^{(k)}$ over k.

4.2 Inverse Propensity-Weighted Maximum Likelihood Estimation

In this section, we show that the federated coefficient estimates from our federated IPW-MLE in Sections 3.2.1 and 3.2.2 have the same asymptotic distributions as those from the combined, individual-level data. Moreover, our federated variance estimator provides a consistent estimator for the variance of the federated coefficients. Before stating the asymptotic results of our federated IPW-MLE, we show a useful lemma that provides the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ on an individual data set.

Lemma 1. *Suppose Assumption 1 holds and we estimate* $e(\mathbf{X}_i)$ *from MLE. Then* $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ *estimated from the weighted log-likelihood function* (4) *is consistent and asymptotically normal,*

$$\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\big) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}\big),$$

where $\boldsymbol{\beta}_0$ *contains the true parameters (see Condition 1),*

$$\mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}} = \mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi}\big(\mathbf{D}_{\boldsymbol{\beta}_0,\varpi} - \mathbf{M}_{\boldsymbol{\beta}_0,\varpi,\boldsymbol{\gamma}}\big)\mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi} \tag{16}$$

with

$$\mathbf{M}_{\boldsymbol{\beta}_0,\varpi,\boldsymbol{\gamma}} = \begin{cases} \mathbf{C}_{\boldsymbol{\beta}_0,\varpi}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}^{\top}_{\boldsymbol{\beta}_0,\varpi} & \textit{ATE weights} \\ \mathbf{C}_{\boldsymbol{\beta}_0,\varpi,1}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}^{\top}_{\boldsymbol{\beta}_0,\varpi,2} + \mathbf{C}_{\boldsymbol{\beta}_0,\varpi,2}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}^{\top}_{\boldsymbol{\beta}_0,\varpi,1} - \mathbf{C}_{\boldsymbol{\beta}_0,\varpi,2}\mathbf{V}_{\boldsymbol{\gamma}}\mathbf{C}^{\top}_{\boldsymbol{\beta}_0,\varpi,2} & \textit{ATT weights.} \end{cases}$$

$\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}$ *is* $\mathbf{A}_{\boldsymbol{\beta},\varpi}$ *evaluated at* $\boldsymbol{\beta}_0$*, with the definition of* $\mathbf{A}_{\boldsymbol{\beta},\varpi}$ *provided in Table 1. Other terms in Eq.* (16) *are defined similarly.* $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}^{-1}_{\boldsymbol{\gamma}_0}$ *if* $e(\mathbf{X}_i, \boldsymbol{\gamma})$ *is correctly specified, and* $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}^{-1}_{\boldsymbol{\gamma}^*}\mathbf{B}_{\boldsymbol{\gamma}^*}\mathbf{A}^{-1}_{\boldsymbol{\gamma}^*}$ *otherwise, where* $\boldsymbol{\gamma}^*$ *minimizes the KL divergence.*

If the true propensity $e(\mathbf{X}_i)$ is used in the weighted log-likelihood function (4)*, then the asymptotic variance simplifies to*

$$\mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle}} = \mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi}\mathbf{D}_{\boldsymbol{\beta}_0,\varpi}\mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi}. \tag{17}$$

Lemma 1 coincides with the results in Wooldridge (2002, 2007) for the case with ATE weights and a correctly specified propensity model. We generalize Wooldridge (2002, 2007) to cases with a misspecified propensity model and with ATT weights. Note that IPW-MLE has the double robustness property, and $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ is a consistent estimator of $\boldsymbol{\beta}_0$ even if the outcome model is misspecified, as long as the propensity model is correctly specified (Wooldridge, 2007).

Since the estimation error of the propensity model affects the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ (Eq. (16) versus Eq. (17)), our federated variance estimator varies with whether the true propensity is used (i.e., Condition 4, see Section 3.2.1 and 3.2.2). Moreover, since $\mathbf{V}_{\boldsymbol{\gamma}}$ in Eq. (16) varies with whether the propensity model is correctly specified (Condition 7), our federated variance estimator also varies with this condition.

Remark 2. An interesting observation from Lemma 1 is that $\mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle}} - \mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}$ is positive semidefinite, implying that $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ is more efficient if the estimated propensity is used in Eq. (4). Hence, even if we know the true propensity score, we suggest using the estimated propensity score (Wooldridge, 2002, Hirano et al., 2003).

We start with stable propensity and outcome models (Conditions 5 and 6). In this case, we use the restricted federated IPW-MLE in Section 3.2.1. When the outcome model is correctly specified (Condition 8), the federated coefficients $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$ have the same asymptotic distribution as $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}}$ estimated from the combined data, and the federated variance $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ is a consistent estimator for the asymptotic variance of $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$.

Theorem 2 (Restricted Federated IPW-MLE for Stable Propensity and Outcome Models)**.** *Suppose Assumption 1 and Conditions 5, 6, and 8 hold. If we use the estimated propensity, we additionally suppose either of the following cases holds true: (a) $\mathbf{V}^{(k),\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}$ is the same across all data sets, or (b) Assumption 2 holds and there exists $M < \infty$ such that $\left\|(\mathbf{A}^{\text{pool}}_{\boldsymbol{\beta}_0,\varpi})^{-1}\mathbf{A}^{(k)}_{\boldsymbol{\beta}_0,\varpi}\right\|_2 \leq M$. As $n_1, \cdots, n_D \to \infty$, we have*

$$n^{1/2}_{\text{pool}}(\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \tag{18}$$

where $\boldsymbol{\beta}_0$ are the true parameters (see Condition 1). If we replace $\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ by $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ and/or replace $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$ by $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}}$, Eq. (18) *continues to hold.*

If we use the true propensity, the above statements continue to hold with $\mathbf{V}^{(k),\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}$, $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$, and $\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ replaced by $\mathbf{V}^{(k),\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle}}$, $\mathbf{V}^{(k),\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle}}$, and $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle}}$, respectively (i.e., the corresponding variance terms for the true propensity).

Federated IPW-MLE improves the efficiency, since the convergence rate is $n^{1/2}_{\text{pool}}$ as shown in Theorem 2, which is faster than $n^{1/2}_k$ on data set k for all k. Similar to Theorem 1, this theorem

allows covariates to have the same or different distributions (Condition 3 either holds or does not). Moreover, Theorem 2 holds regardless of whether we use the true or estimated propensity (Condition 4 either holds or does not). In practice, even if IPW-MLE uses the estimated propensity, we could use the federated variance estimator that works for the true propensity case, which takes a much simpler form (see Remark 2), but overestimates the asymptotic variance.

Next, when the propensity or outcome model is unstable (Conditions 5 and/or 6 are violated), we specify a more flexible model for the combined data that adjusts for individual data sets and use the federated IPW-MLE in Section 3.2.2. The following proposition shows that Theorem 2 continues to hold for the federated IPW-MLE for the unstable propensity or outcome model.

Proposition 4 (Unrestricted Federated IPW-MLE for Unstable Propensity and/or Outcome Models)**.** *Suppose Assumptions 1 and 2 hold, Conditions 3 and/or 6 are violated, and* $\mathbf{A}^{\text{fed}}_{\boldsymbol{\beta}_{0,\varpi}}$ *and* $\mathbf{A}^{\text{fed}}_{\boldsymbol{\gamma}_0}$ *(or* $\mathbf{A}^{\text{fed}}_{\boldsymbol{\gamma}^*}$ *for misspecified propensity models) are full rank. If we use the method in Section 3.2.2 to federate coefficients and variance, then Theorem 2 continues to hold.*

4.3 AIPW Estimation

Our federated AIPW estimators in Sections 3.3.1 and 3.3.2 are based on the asymptotic linear property of the AIPW estimator (Robins et al., 1994, Tsiatis and Davidian, 2007). For completeness, we state this property in the following lemma.

Lemma 2 (Adapted from Tsiatis and Davidian (2007) and Chernozhukov et al. (2017))**.** *Suppose at least one condition holds: (a)* $\mu_{(w)}(\mathbf{x}, \boldsymbol{\beta}_{0,w})$ *correctly specifies* $\mathbb{E}[Y_i|\mathbf{X}_i, W_i = w]$ *for some* $\boldsymbol{\beta}_{0,w}$ *and* $w \in \{0,1\}$*, or (b)* $e(\mathbf{X}_i, \boldsymbol{\gamma}_0)$ *correctly specifies* $\text{pr}(W_i = 1|\mathbf{X}_i)$ *for some* $\boldsymbol{\gamma}_0$*. Then the AIPW estimator* $\hat{\tau}_{\text{aipw}}$ *satisfies*

$$\sqrt{n}(\hat{\tau}_{\text{aipw}} - \tau_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(\mathbf{X}_i, W_i, Y_i) + o_p(1) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{V}_\tau\big), \tag{19}$$

for the influence function $\phi(\mathbf{X}_i, W_i, Y_i)$ *with* $\mathbb{E}[\phi(\mathbf{X}_i, W_i, Y_i)] = 0$ *and* $\mathbf{V}_\tau = \mathbb{E}[\phi(\mathbf{X}_i, W_i, Y_i)^2]$*. If the estimand is ATE,*

$$\begin{aligned}\phi(\mathbf{X}_i, W_i, Y_i) =& \mu_{(1)}(\mathbf{X}_i, \boldsymbol{\beta}_1^*) - \mu_{(0)}(\mathbf{X}_i, \boldsymbol{\beta}_0^*) + \frac{W_i}{e(\mathbf{X}_i, \gamma^*)}(Y_i - \mu_{(1)}(\mathbf{X}_i, \boldsymbol{\beta}_1^*)) \\ &- \frac{(1 - W_i)}{1 - e(\mathbf{X}_i, \gamma^*)}(Y_i - \mu_{(0)}(\mathbf{X}_i, \boldsymbol{\beta}_0^*)) - \tau_0,\end{aligned} \tag{20}$$

and if the estimand is ATT,

$$\phi(\mathbf{X}_i, W_i, Y_i) = W_i\big(Y_i - \mu_{(1)}(\mathbf{X}_i, \boldsymbol{\beta}_1^*)\big) - \frac{e(\mathbf{X}_i, \gamma^*)(1 - W_i)}{1 - e(\mathbf{X}_i, \gamma^*)}\big(Y_i - \mu_{(0)}(\mathbf{X}_i, \boldsymbol{\beta}_0^*)\big) - \tau_0, \tag{21}$$

where $\boldsymbol{\gamma}^* = \lim_{n\to\infty} \hat{\boldsymbol{\gamma}}$*,* $\boldsymbol{\beta}_w^* = \lim_{n\to\infty} \hat{\boldsymbol{\beta}}_w$*, and at least one equality holds: (a)* $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$*, or (b)* $\boldsymbol{\beta}_w^* = \boldsymbol{\beta}_{0,w}$ *for* $w \in \{0,1\}$*.*

We can see from Lemma 2 that the score $\hat{\phi}(\mathbf{X}_i, W_i, Y_i)$ in the definition of $\hat{\tau}_{\text{aipw}}$ is an estimator of $\tau_0 + \phi(\mathbf{X}_i, W_i, Y_i)$ (recall Section 2.2.3). Lemma 2 formally states the doubly robust property we mentioned in Section 2.2.3: $\hat{\tau}_{\text{aipw}}$ continues to be consistent and asymptotically normal if either $\boldsymbol{\gamma}^* \neq \boldsymbol{\gamma}_0$ (i.e., misspecified propensity model) or $\boldsymbol{\beta}_w^* \neq \boldsymbol{\beta}_{0,w}$ (i.e., misspecified outcome model), but not both. The following theorem shows that (a) our federated AIPW estimator $\hat{\tau}_{\text{aipw}}^{\text{fed}}$ has the same asymptotic distribution as the AIPW estimator $\hat{\tau}_{\text{aipw}}^{\text{pool}}$ from the combined data, and (b) our federated variance estimator $\hat{\mathbf{V}}_\tau^{\text{fed}}$ is consistent.

Theorem 3. *Suppose either of the following cases holds: (a) $\phi(\mathbf{X}_i, W_i, Y_i)$ is the same for all data sets, and we use the federation formulas in Section 3.3.1; or (b) Assumption 2 holds, $\phi(\mathbf{X}_i, W_i, Y_i)$ varies with the data set, and we use the federation formulas in Section 3.3.2. As $n_1, \cdots, n_D \to \infty$, we have*

$$n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_\tau^{\text{pool}})^{-1/2}(\hat{\tau}_{\text{aipw}}^{\text{fed}} - \tau_0) \xrightarrow{d} \mathcal{N}(0,1), \tag{22}$$

where $\tau_0 = \sum_{k=1}^{D} p_k \tau_0^{(k)}$ is the weighted ATE or ATT. If we replace $\hat{\mathbf{V}}_\tau^{\text{pool}}$ by $\hat{\mathbf{V}}_\tau^{\text{fed}}$ and/or replace $\hat{\tau}_{\text{aipw}}^{\text{fed}}$ by $\hat{\tau}_{\text{aipw}}^{\text{pool}}$, Eq. (22) continues to hold.

For stable propensity and outcome models (i.e., when Conditions 5 and 6 hold), $\tau_0^{(k)}$ equals τ_0 for all k. Since $\hat{\tau}_{\text{aipw}}^{\text{fed}}$ has the convergence rate $n_{\text{pool}}^{1/2}$, we can improve the efficiency by federation. When at least one of the propensity or outcome models is unstable (i.e., either Condition 5 or 6 is violated), $\tau_0^{(k)}$ may be different for different k. Then, τ_0 defined on the combined data is a weighted average of $\tau_0^{(k)}$ by the relative sizes of data sets (i.e., $p_k = \lim n_k/n_{\text{pool}}$). In this case, our federated AIPW estimator in Section 3.3.2 provides a consistent and asymptotic normal estimator for τ_0.

5 Empirical Studies Based on Medical Claims Data

In this section, we apply our methods to conduct retrospective analyses on whether patients exposed to alpha blockers (a class of drugs including tamsulosin and doxazosin; broadly known as α_1-AR antagonists), as compared to unexposed patients, have a reduced risk of adverse outcomes (e.g., mechanical ventilation and death) in lower respiratory tract infections. Lower respiratory tract infections are any infections in the lungs or below the voice box, including pneumonia and acute respiratory distress (ARD). The effectiveness of alpha blockers is of particular interest as no human clinical trials have studied this question despite promising evidence in mice (Staedtke et al., 2018). Our analysis builds on the studies by Konig et al. (2020), Koenecke et al. (2021), Rose et al. (2021), Powell et al. (2021), and Thomsen et al. (2021). Following Koenecke et al. (2021), we have access to patient records in two de-identified databases that cannot legally or ethically be combined together: the IBM MarketScan®Research Database (which we refer to as MarketScan) and Optum's Clinformatics®Data Mart Database, a commercial and Medicare Advantage claims database (which we refer to as Optum). We seek to make an inference about the effect of alpha

blockers using these two databases. To do so, we require federated methods that can obtain the point and variance estimates using only summary-level information on individual databases.

Our federated methods may, under different conditions, produce different estimates from other methods such as inverse variance weighting (IVW). Researchers hence face the dilemma of deciding which estimates are more trustworthy, as determined by what estimates are closer to what would have been obtained from the combined, individual-level data.

In this section, we suggest a general federated analysis pipeline and use the study of alpha blockers as an illustrative example. First, we construct subsamples from one cohort (e.g., within $\mathcal{C}_{\text{M,ARD}}$ for ARD patients identified in the Marketscan database) to mimic the demographics of the underlying databases and validate various federated methods. Since we are able to combine the subsamples constructed from one cohort, we compare the federated estimates from various methods with the estimates from the combined data, as shown in Section 5.2. Second, we use the best federated methods from the first step, which are our unrestricted federated MLE and unrestricted federated IPW-MLE, to combine the summary-level information in MarketScan and Optum in Section 5.3. We expect the performance of these federated methods to carry over to the federation of MarketScan and Optum given their similar demographics. The federated methods we apply in the second step are robust to model misspecification and instability, supported by the asymptotic theory, and robust across simulation studies conducted in the first step (see Appendix A).[12]

5.1 Study Definitions

We follow the study definitions in Koenecke et al. (2021).

Participants We study two cohorts of patients who were diagnostically coded in U.S. hospitals with acute respiratory distress (ARD) from each of the MarketScan and Optum databases. We further study two cohorts of patients diagnostically coded in U.S. hospitals with pneumonia from each of the MarketScan and Optum databases.

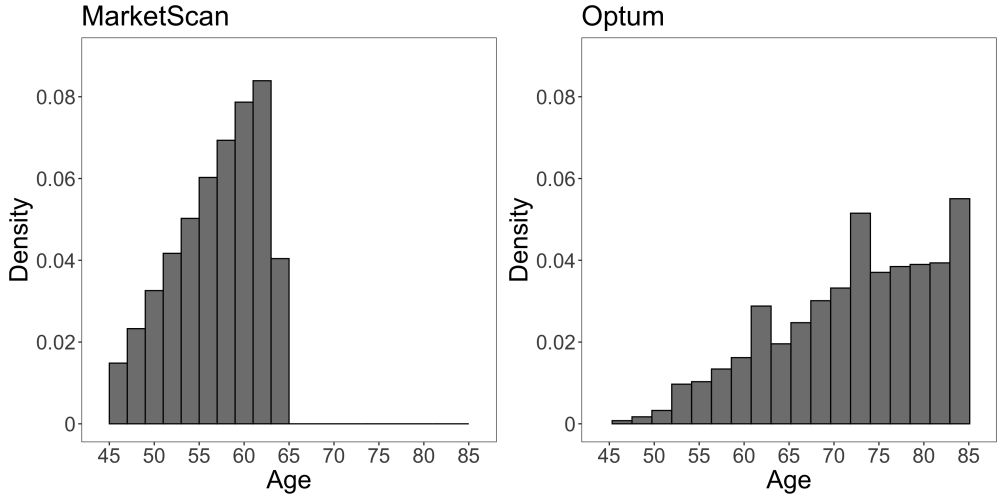
We limit the study to older men because alpha blockers are widely used as a treatment in the U.S. for benign prostatic hyperplasia (BPH), a common condition in older men that is clinically unrelated to the respiratory system. More specifically, we focus on men over the age of 45 so that a large portion of the exposed group faces similar risks of poor outcomes from respiratory conditions as the unexposed group, thus mitigating confounding by indication.[13] In addition, we

[12]Note that our findings reproduce similar results to Koenecke et al. (2021), validating the prior result suggesting that alpha blockers are effective in reducing ventilation and death in ARD and pneumonia patients; however, the numbers differ slightly because our federated methods presented here are improved from those used in Koenecke et al. (2021). Koenecke et al. (2021) only use the treatment coefficient and variance to federate estimators, whereas here we use the full variance-covariance matrix from all covariates, which is a more robust approach for heterogeneous data.

[13]Note that this limits our analysis' validity to older men due to their being the dominant population historically being prescribed alpha blockers. However, we recognize the importance of studying other demographics, such as women and younger men, in clinical studies (Holdcroft, 2007, McMurray et al., 1991); extrapolating our results to these demographics would require additional assumptions as noted in (Powell et al., 2021).

enforce a maximum age of 85 years to reflect the ongoing clinical trials investigating prazosin (an alpha blocker) and its effects on COVID-19 patients.[14]

Figure 7: Histograms of Patient Age in MarketScan and Optum



We restrict all patients in both MarketScan and Optum databases to be over the age of 45. While patient data from the MarketScan database only include patients younger than age 65, a majority of the patients in the Optum database are over 65 years old.

After the restrictions on sex and age, we obtain a cohort of 12,463 ARD inpatients and a cohort of 103,681 pneumonia inpatients from the MarketScan database (denoted as $\mathcal{C}_{\text{M,ARD}}$ and $\mathcal{C}_{\text{M,PNA}}$, respectively), and a cohort of 6,084 ARD inpatients and a cohort of 234,993 pneumonia inpatients from the Optum database (denoted as $\mathcal{C}_{\text{O,ARD}}$ and $\mathcal{C}_{\text{O,PNA}}$, respectively).

The demographics of patients in the MarketScan and Optum databases differ in two aspects. First, Optum includes older patients as MarketScan only includes patients up to age 65 due to Medicare exclusions (see Figure 7 for the distribution of patient age on MarketScan and Optum). Second, Optum has more recent patient records from the fiscal year 2004 to 2019, while MarketScan only has patient records from the fiscal year 2004 to 2016.

Treatment W_i Treatment W_i is binary (either 1 or 0) and indicates whether a patient is exposed to alpha blockers ($W_i = 1$) or not ($W_i = 0$).[15]

Outcome Y_i Outcome Y_i is binary (either 1 or 0) and indicates whether a patient both received mechanical ventilation as a medical procedure and then had an in-hospital death (corresponding to $Y_i = 1$).[16]

[14] See `https://clinicaltrials.gov/ct2/show/NCT04365257`.

[15] Patients are considered exposed if they were prescribed α_1-AR antagonists (doxazosin, alfuzosin, prazosin, silodosin, terazosin, or tamsulosin) for at least six of the 12 months prior to their first inpatient admission date for either ARD or pneumonia.

[16] Death outcomes were identified by in-hospital deaths from MarketScan claims records, or by patient month of death within 1 month of relevant ARD or pneumonia inpatient visit month from the Optum claims records.

Potential Confounders $\mathbf{X}_i$ $\mathbf{X}_i$ consists of age, fiscal year, and health-related confounders.[17]

Model Specification We estimate whether the exposure to alpha blockers has a reduced risk of adverse outcomes on each cohort from the following outcome model specification:

$$\frac{f(Y_i = 1|\mathbf{X}_i, W_i)}{f(Y_i = 0|\mathbf{X}_i, W_i)} = W_i \beta_{0,w}^{(A_i)} + \mathbf{X}_i^\top \boldsymbol{\beta}_{0,\mathbf{X}}^{(A_i)},$$

where $A_i \in \{\mathrm{M}, \mathrm{O}\}$ denotes the database to which patient i belongs (either MarketScan or Optum). We estimate $\beta_{0,w}^{(A_i)}$ and $\boldsymbol{\beta}_{0,\mathbf{X}}^{(A_i)}$ from MLE and IPW-MLE. The value $\exp(\beta_{0,w}^{(A_i)})$ is the odds ratio for alpha blockers on patients being ventilated and dying; the odds ratio metric is commonly reported in medical studies (Szumilas, 2010) and represents the odds that Y_i will occur given treatment exposure, compared to the odds that Y_i will occur without treatment. The propensity scores used in IPW-MLE are estimated from the following propensity model specification:

$$\frac{\mathrm{pr}(W_i = 1|\mathbf{X}_i)}{\mathrm{pr}(W_i = 0|\mathbf{X}_i)} = \mathbf{X}_i^\top \boldsymbol{\gamma}_0^{(A_i)}.$$

In Sections 5.2 and 5.3, we focus on the federated analysis based on MLE and IPW-MLE that produce estimates of odds ratios similar to Koenecke et al. (2021). We extend our alpha blocker analysis based on federated AIPW in Appendix B, but note that results are not directly comparable or as easily interpretable as the odds ratios arising from our federated IPW-MLE method because AIPW produces a different measure based on ATE or ATT.

5.2 Federated Results Based on Sampling from One Medical Claims Data Set

In this section, we construct subsamples from one cohort to reflect patient demographics from the medical claims data; we then compare estimates from various federated methods using the combined data. Our purpose in this step is to observe how well the federated methods recover the known result from the combined data in a setting where combining data is permissible. We can then select the most effective federated methods and apply these methods to combine the summary-level information from MarketScan and Optum in Section 5.3. Recall that combining individual-level raw data from these two sources is not permissible.

We start by presenting our approach to construct subsamples from one cohort (i.e., one of $\mathcal{C}_{\mathrm{M,ARD}}$, $\mathcal{C}_{\mathrm{M,PNA}}$, or $\mathcal{C}_{\mathrm{O,PNA}}$) in Section 5.2.1. We exclude $\mathcal{C}_{\mathrm{O,ARD}}$ in this analysis as $\mathcal{C}_{\mathrm{O,ARD}}$ has a very small sample size, yielding singularity issues during computation. Then, in Section 5.2.2 we list the tested methods for combining point and variance estimates across subsamples, including our federated methods and inverse variance weighting (IVW), and list benchmark estimation methods.

[17] Health-related confounders include total weeks with inpatient admissions in the prior year, total outpatient visits in the prior year, total days as an inpatient in the prior year, total weeks with inpatient admissions in the prior two months, and comorbidities identified from healthcare encounters in the prior year: hypertension, ischemic heart disease, acute myocardial infarction, heart failure, chronic obstructive pulmonary disease, diabetes mellitus, and cancer.

We compare results from federated methods against the benchmark methods in Section 5.2.3 and show that our unrestricted federated MLE and unrestricted federated IPW-MLE are the most effective federated methods.

5.2.1 Sampling Schemes for Subsamples

In the base case, we construct two equally-sized subsamples, denoted as $\mathcal{S}_1$ and $\mathcal{S}_2$, based on the patient records from one cohort (denoted as $\mathcal{C}$). One subsample is constructed to have more elderly patients than the other to reflect the age heterogeneity between MarketScan and Optum.

To construct these two subsamples, we first split the cohort $\mathcal{C}$ into two equally-sized sub-cohorts, denoted as $\mathcal{C}_1$ and $\mathcal{C}_2$, by age (such that $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ and $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$). $\mathcal{C}_1$ only has records of the patients whose age is below the median of all patients in $\mathcal{C}$, and $\mathcal{C}_2$ is the complement of $\mathcal{C}_1$. Without loss of generality, suppose we construct $\mathcal{S}_2$ to have more elderly patients than $\mathcal{S}_1$. Then, we sample 80% of the patient records in $\mathcal{S}_1$ from $\mathcal{C}_1$ with replacement, with the remaining 20% sampled from $\mathcal{S}_2$ with replacement. Separately, 20% of the patient records in $\mathcal{S}_2$ are sampled from $\mathcal{C}_1$ with replacement, with the remaining 80% sampled from $\mathcal{S}_2$ with replacement.

As a robustness check, we consider other approaches in Appendix B to construct subsamples, including varying the sampling ratios from $\mathcal{C}_1$ and $\mathcal{C}_2$ to construct subsamples, constructing sub-cohorts using different heterogeneous attributes (e.g., varying with both age and fiscal year), varying subsample sizes, and varying the number of subsamples. In addition, we conduct data-driven simulations in Appendix C, where we construct subsamples by sampling patients' covariates from $\mathcal{C}$, and simulating treatment and outcome according to some known propensity and outcome models. We can then control for whether the propensity and outcome models are correctly specified and stable across subsamples in these data-driven simulations. The results in Appendices B and C show that our findings in Section 5.2.3 are consistent with our other approaches in constructing subsamples.

5.2.2 Estimation and Benchmarks

In this subsection, we list the benchmark and federated estimation approaches for the treatment coefficient in the outcome model and its variance.

Restricted Benchmarks In this case, we assume parameters in the propensity and outcome models are stable across subsamples. On the combined data, we specify the propensity and outcome models as

$$\begin{aligned} \frac{\text{pr}(W_i = 1|\mathbf{X}_i)}{\text{pr}(W_i = 0|\mathbf{X}_i)} =& \mathbf{X}_i^\top \boldsymbol{\gamma}_0 \\ \frac{f(Y_i = 1|\mathbf{X}_i, W_i)}{f(Y_i = 0|\mathbf{X}_i, W_i)} =& W_i \beta_{0,w} + \mathbf{X}_i^\top \boldsymbol{\beta}_{0,\mathbf{X}}. \end{aligned} \tag{23}$$

The restricted benchmarks are set as the estimated $\beta_{0,w}$ in (23) and its estimated variance from the combined data, denoted as $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}}$, respectively.

Unrestricted Benchmarks In this case, we assume the parameters of covariates $\mathbf{X}_i$ in the propensity and outcome models are unstable across subsamples. On the combined data, we specify a flexible functional form for the propensity and outcome models

$$\begin{aligned}\frac{\text{pr}(W_i=1|\mathbf{X}_i)}{\text{pr}(W_i=0|\mathbf{X}_i)} =&\mathbf{X}_i^\top\left(\mathbf{1}(A_i=1)\boldsymbol{\gamma}_0^{(1)}+\cdots+\mathbf{1}(A_i=D)\boldsymbol{\gamma}_0^{(D)}\right)\\ \frac{f(Y_i=1|\mathbf{X}_i,W_i)}{f(Y_i=0|\mathbf{X}_i,W_i)} =&W_i\beta_{0,w}+\mathbf{X}_i^\top\left(\mathbf{1}(A_i=1)\boldsymbol{\beta}_{0,\mathbf{X}}^{(1)}+\cdots+\mathbf{1}(A_i=D)\boldsymbol{\beta}_{0,\mathbf{X}}^{(D)}\right),\end{aligned} \tag{24}$$

where $A_i \in \{1,\cdots,D\}$ indicates to which subsample the patient record $(\mathbf{X}_i, W_i, Y_i)$ belongs (in our results presented in this section, $D=2$, but we vary the value of D in robustness checks in Appendix B). Note that $\beta_{0,w}$ is stable across subsamples, which can be interpreted as the average treatment coefficient across subsamples. The unrestricted benchmarks are set as the estimated $\beta_{0,w}$ in (24) and its estimated variance from the combined data, denoted as $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{unr}}_{\beta_w,\text{bm}}$, respectively.

Restricted Federated Estimators We estimate $\boldsymbol{\gamma}_0, \beta_{0,w}$, and $\boldsymbol{\beta}_{0,\mathbf{X}}$ on each subsample (denoted as $\hat{\boldsymbol{\gamma}}^{(j)}, \hat{\beta}_w^{(j)}$, and $\hat{\boldsymbol{\beta}}_{\mathbf{X}}^{(j)}$). Under the model specification (23) on the combined data, we use our restricted federated MLE and IPW-MLE to combine $\hat{\boldsymbol{\gamma}}^{(j)}, \hat{\beta}_w^{(j)}$, and $\hat{\boldsymbol{\beta}}_{\mathbf{X}}^{(j)}$ for all j. Denote the restricted federated MLE for $\beta_{0,w}$ and its variance as $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ and $\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{mle}}$, respectively. Denote the restricted federated IPW-MLE for $\beta_{0,w}$ and its variance as $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ and $\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{ipw-mle}}$, respectively.

Unrestricted Federated Estimators We estimate $\boldsymbol{\gamma}_0$, $\beta_{0,w}$, and $\boldsymbol{\beta}_{0,\mathbf{X}}$ on each subsample (denoted as $\hat{\boldsymbol{\gamma}}^{(j)}$, $\hat{\beta}_w^{(j)}$, and $\hat{\boldsymbol{\beta}}_{\mathbf{X}}^{(j)}$). Under the flexible model specification (24) on the combined data, we use our unrestricted federated MLE and IPW-MLE to combine $\hat{\boldsymbol{\gamma}}^{(j)}$, $\hat{\beta}_w^{(j)}$, and $\hat{\boldsymbol{\beta}}_{\mathbf{X}}^{(j)}$ for all j. We use $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$, $\hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{mle}}$, $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$, and $\hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{ipw-mle}}$ to denote the unrestricted federated MLE and IPW-MLE.

Inverse Variance Weighting (IVW) IVW is commonly used in meta-analysis (DerSimonian and Laird, 1986, Whitehead and Whitehead, 1991, Sutton and Higgins, 2008, Hartung et al., 2011), and it is an appropriate method for (equally weighted) MLE assuming the outcome model is correctly specified and stable across data sets.[18] Specifically, we estimate $\boldsymbol{\gamma}_0$, $\beta_{0,w}$, and $\boldsymbol{\beta}_{0,\mathbf{X}}$ on each subsample. Under the restricted model specification (23) on the combined data, we combine $\hat{\beta}_w^{(j)}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{X}}^{(j)}$ for all j together weighted by their inverse variance $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{(j)}$ (see Section 2.4.3). Denote the inverse variance weighting estimator for $\beta_{0,w}$ and its variance as $\hat{\beta}_{w,\text{ivw}}$ and $\hat{V}_{\beta_w,\text{ivw}}$, respectively.

5.2.3 Results

In this section, we compare our restricted and unrestricted federated estimators (MLE/IPW-MLE) and IVW with the benchmarks. We refer to the accuracy of a federated estimator as its difference

[18] Under correct model specification, the information matrix equality holds and the Hessian matrix asymptotically equals the inverse variance matrix, so Hessian weighting is the same as inverse variance weighting.

from the benchmark. The benchmark (i.e., $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}}$, $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, and $\hat{V}^{\mathbf{unr}}_{\beta_w,\text{bm}}$) is the result we would obtain using the combined data, so this is the standard by which we evaluate each federated estimator.

We demonstrate three main findings in this section:

1. Our restricted and unrestricted federated MLE and federated IPW-MLE are more accurate than IVW, regardless of whether we assume stable propensity and outcome models or not.

2. Federated MLE and federated IPW-MLE yield similar point and variance estimates, alleviating our concern about outcome model misspecification as IPW-MLE is doubly robust (see Section 5.3 for further discussion).

3. Our unrestricted federated MLE and federated IPW-MLE are preferable to the restricted methods for robustness to model instability and for more accurate variance estimates.

Comparison Between Federated MLE and IVW We compare our federated MLE (restricted: $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ and $\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{ipw-mle}}$; unrestricted: $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ and $\hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{ipw-mle}}$) and inverse variance weighting ($\hat{\beta}_{w,\text{ivw}}$ and $\hat{V}_{\beta_w,\text{ivw}}$) to the benchmarks (restricted: $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}}$; unrestricted: $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{unr}}_{\beta_w,\text{bm}}$) in Table 5. We use the sandwich formula for $\mathbf{V}_{\boldsymbol{\beta}}$ to allow for the violation of information matrix equality in finite samples.[19] Assuming the correct outcome model specification, the Mean Absolute Error (MAE) of our restricted federated MLE and IVW point estimates are 0.0449 and 0.0724, respectively, on $\mathcal{C}_{\text{M,ARD}}$. Comparably, the MAE of our restricted federated MLE point estimate is smaller than that of IVW on the larger data sets $\mathcal{C}_{\text{M,PNA}}$ and $\mathcal{C}_{\text{O,PNA}}$. Similar findings carry over to the variance estimates under the stable outcome model assumption, and carry over to both the point and variance estimates allowing for outcome model instability (unrestricted benchmarks) as shown in Table 5. As such, our restricted and unrestricted federated MLE point estimates are more accurate than IVW (even when the outcome model is stable and correctly specified, as the information matrix equality does not hold in finite samples), though accuracy gains with federated MLE may decrease with sample size.

Comparison Between Federated IPW-MLE and IVW The implementations of our federated IPW-MLE always differ from IVW as shown in Section 3.[20] We compare our federated IPW-MLE (restricted: $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ and $\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{ipw-mle}}$; unrestricted: $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ and $\hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{ipw-mle}}$) and inverse variance weighting (IVW: $\hat{\beta}_{w,\text{ivw}}$ and $\hat{V}_{\beta_w,\text{ivw}}$) to the benchmarks (restricted: $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}}$; unrestricted: $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{unr}}_{\beta_w,\text{bm}}$) in Table 5. Assuming stable propensity and outcome models, the MAE of our restricted federated IPW-MLE and IVW point estimates are 0.0526 and 0.9128, respectively, on $\mathcal{C}_{\text{M,ARD}}$. Comparably, the MAE of our restricted federated IPW-MLE point

[19] The information equality could be violated when the outcome model is misspecified or we have a small sample size.

[20] We can show that IVW is consistent and therefore appropriate in the special case where subsamples have the same populations (more specifically, where the covariate distribution is stable, and the propensity and outcome models are stable). In general cases, IVW is inconsistent.

Table 5: Comparison Between Restricted/Unrestricted Federated Estimators and IVW with Corresponding Restricted/Unrestricted Benchmarks

(a) MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
MS ARD	*-0.7114*	0.0724	0.0449	0.0463
MS PNA	*-0.2348*	0.0132	0.0073	0.0088
Optum PNA	*-0.1545*	0.0037	0.0011	0.0015
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
MS ARD	*0.0852*	0.0196	0.0114	0.0074
MS PNA	*0.0173*	0.0010	0.0006	0.0002
Optum PNA	*0.0026*	0.0000	0.0000	0.0000

(b) MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
MS ARD	*-0.7097*	0.0733	0.0465	0.0432
MS PNA	*-0.2330*	0.0131	0.0070	0.0066
Optum PNA	*-0.1548*	0.0038	0.0019	0.0009
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
MS ARD	*0.0851*	0.0198	0.0113	0.0072
MS PNA	*0.0173*	0.0010	0.0006	0.0002
Optum PNA	*0.0026*	0.0000	0.0000	0.0000

(c) IPW-MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
MS ARD	*-0.7096*	0.9128	0.0526	0.0780
MS PNA	*-0.3019*	0.3883	0.0094	0.0115
Optum PNA	*-0.1832*	0.0536	0.0011	0.0043
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
MS ARD	*0.0966*	0.0690	0.0282	0.0192
MS PNA	*0.0244*	0.0103	0.0024	0.0006
Optum PNA	*0.0031*	0.0003	0.0001	0.0000

(d) IPW-MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
MS ARD	*-0.7495*	0.8728	0.1089	0.0302
MS PNA	*-0.3034*	0.3869	0.0144	0.0027
Optum PNA	*-0.1852*	0.0517	0.0043	0.0002
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
MS ARD	*0.0835*	0.0559	0.0159	0.0054
MS PNA	*0.0242*	0.0102	0.0022	0.0002
Optum PNA	*0.0031*	0.0003	0.0001	0.0000

These tables compare restricted and unrestricted federated MLE and federated IPW-MLE (restricted: $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$,$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ and $\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{mle}}$,$\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{ipw-mle}}$; unrestricted: $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$,$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ and $\hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{mle}}$,$\hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{ipw-mle}}$) estimators, and inverse variance weighted ($\hat{\beta}_{w,\text{ivw}}$ and $\hat{V}_{\beta_w,\text{ivw}}$) estimators to the restricted and unrestricted benchmarks (restricted: $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}}$; unrestricted: $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{unr}}_{\beta_w,\text{bm}}$). In the federated MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\beta}}$ which is appropriate when the outcome model is misspecified. In the federated IPW-MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\gamma}}$ which allows for propensity model misspecification (i.e., when Condition 7 is violated). We construct subsamples from the MarketScan ARD (MS ARD) cohort, and from the MarketScan and Optum pneumonia (MS PNA and Optum PNA) cohorts, letting $D = 2$. For MS ARD, $n_1 = n_2 = n_{\text{pool}}/2 = 6,000$; for MS PNA and Optum PNA, $n_1 = n_2 = n_{\text{pool}}/2 = 30,000$. We use ATE weighting in IPW-MLE in these tables. We report the results with ATT weighting in Table 6 in Appendix B; the results with ATT weighting are close to the results with ATE weighting. The mean absolute error (MAE) is calculated relative to the benchmark mean values (first column of each table) based on 50 iterations of independent sampling of subsamples. Note that the combined data $\mathcal{C}_1 \cup \mathcal{C}_2$ vary across iterations, so we report the average of the benchmarks ($\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{bm}}$) across iterations, yielding slight noise in the benchmark means reported.

estimate is much smaller than that of IVW on the larger data set $\mathcal{C}_{\text{M,PNA}}$, to a similar extent as on $\mathcal{C}_{\text{M,ARD}}$ (with respect to relative MAE compared with the mean of $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$). Similar findings carry over to the variance estimates under the stable outcome model assumption, and carry over to both the point and variance estimates when unstable propensity and outcome models are allowed (unrestricted benchmarks), as shown in Table 5. We conclude that our federated IPW-MLE is much more accurate than inverse variance weighting for both point and variance estimates.

Comparison Between Federated MLE and Federated IPW-MLE The means of the restricted benchmarks, $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}}$, are -0.7114 and 0.0852, respectively for MLE on $\mathcal{C}_{\text{M,ARD}}$. These values are very close to the means of the IPW-MLE benchmarks, -0.7096 and 0.0966, respectively (see Table 5). Similar findings carry over to the unrestricted benchmarks. As IPW-MLE is doubly robust, the similar point and variance estimates from MLE and IPW-MLE alleviates our concern of a biased estimate due to outcome model misspecification.

Comparison Between Restricted and Unrestricted Federated Estimators Our unrestricted federated MLE and federated IPW-MLE estimators specify flexible propensity and outcome models that adjust for data sets (e.g., in Eq. (24)). When the true models are stable, the unrestricted federated estimators are consistent, but they could be less efficient than the restricted federated estimators. As shown in Table 5, compared to the restricted benchmarks, the MAE values of $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ and $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ are generally smaller than the MAE values of $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ and $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$, which is consistent with the efficiency loss of the unrestricted federated estimators when the true model is stable. On the other hand, when the true model is unstable, the restricted federated estimators could be biased. As again shown in Table 5, compared to the unrestricted benchmarks, the MAE values of $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ and $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ are much larger than the MAE values of $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ and $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$. Interestingly, we consistently observe that the unrestricted variance estimates have a smaller MAE than the restricted variance estimates, compared to either the restricted benchmarks or unrestricted benchmarks. This observation is because of the empirical finding that $\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{mle}} \underset{①}{\leq} \hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{mle}} \underset{②}{\leq} \hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}} \underset{③}{\leq} \hat{V}^{\mathbf{unr}}_{\beta_w,\text{bm}}$ for MLE (and the same ordering for IPW-MLE), as shown via simulations in Table 12 of Appendix C. Unrestricted estimators with a flexible model specification are generally less efficient than restricted estimators (see Proposition 1), leading to Inequalities ① and ③. Moreover, we observe that both restricted and unrestricted federated estimators generally underestimate the true variance in finite samples (even though both are consistent), leading to $\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{mle}} < \hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}}$. The underestimated amount $\hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}} - \hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{mle}}$ is generally larger than the efficiency loss $\hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{mle}} - \hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{mle}}$, leading to Inequality ②.

5.3 Federation Across Two Medical Claim Data Sets

In Section 5.2.3, we demonstrate that our unrestricted federated MLE and unrestricted federated IPW-MLE perform best in a single data set where a pooled data estimate was obtainable. Now we show the application of these methods on two data sets, MarketScan and Optum, which cannot be pooled at the level of individual patients. To start, we present the results on individual cohorts.

Cohort-Specific Results The coefficient on alpha blockers is consistently negative on the individual cohorts of ARD patients ($\mathcal{C}_{\text{M,ARD}}$ and $\mathcal{C}_{\text{O,ARD}}$) and pneumonia patients ($\mathcal{C}_{\text{M,PNA}}$ and $\mathcal{C}_{\text{O,PNA}}$) as shown in Figure 8, implying a reduced risk of adverse outcomes for ARD and pneumonia patients who were exposed to alpha blockers.

We additionally note that some coefficients (though, none with statistical significance) are

of different magnitudes or signs in the outcome model across the two databases. For example, coefficients for patient age and weeks with prior inpatient admissions (i.e., a metric for overall health) have different signs between MarketScan and Optum results in the ARD cohorts (see Figures 10a and 10b in Appendix B).[21] This raises three potential concerns: model instability, model misspecification, and unobserved confounders across the two databases, which we ameliorate as follows.

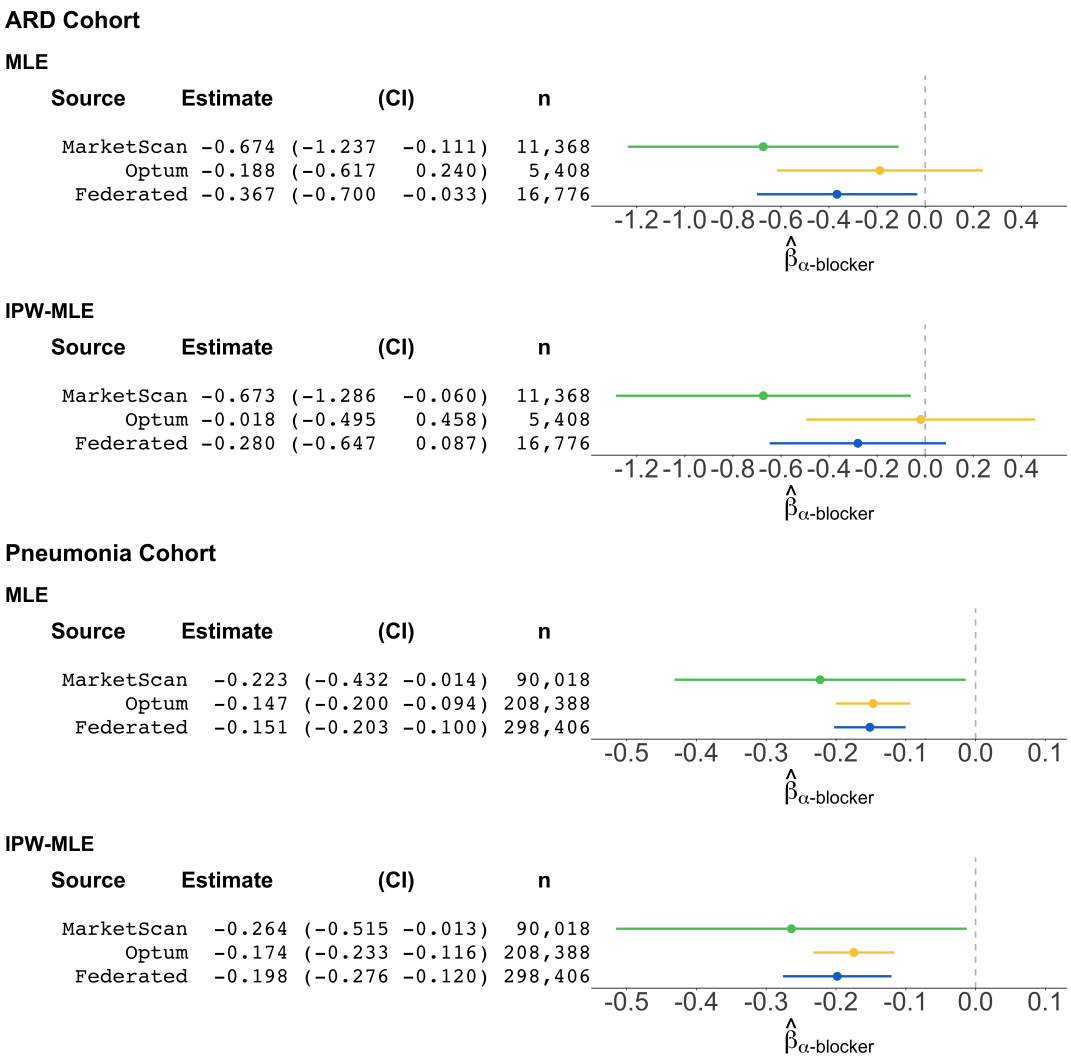
First, model instability could be due to the different populations underlying these two databases as shown in Figure 7, as well as the heterogeneous response of outcomes to the treatment and confounders. We consider a flexible functional form for the combined data that adjusts for data sets (see Eq. (24)), and our unrestricted federated MLE or unrestricted federated IPW-MLE is appropriate in the presence of model instability. Second, model misspecification could exist if the response is indeed the same across two databases, but there exists a coefficient difference in the estimated outcome models. To protect against this possibility, we suggest using IPW-MLE due to its doubly robust properties (as opposed to MLE), which allows for consistency if the propensity model is correctly specified. Third, we may have unobserved confounders in the outcome model. However, we have largely controlled for unobserved confounders in our approach to construct the cohorts (as discussed in the **Participants** paragraph in Section 5.1).[22]

Federated Results Figure 8 shows the point estimates and confidence intervals from MarketScan and Optum as well as from our unrestricted federated MLE and unrestricted federated IPW-MLE estimators. As desired, the federated estimates of the effect of alpha blockers lie between the estimates on MarketScan and Optum for both ARD and pneumonia patients, and they approximate the average effect of alpha blockers on all ARD or pneumonia patients across two databases (recall the estimates from IVW may not lie between those on MarketScan and Optum as shown in Figure 1 and Figure 9 in Appendix B). The federated confidence intervals are generally narrower than those on individual databases, implying more statistical power from federation to detect the effect of alpha blockers. As a robustness check, we report the average treatment effect of alpha blockers on all patients (i.e., ATE) and on treated patients (i.e., ATT) from AIPW on MarketScan and Optum, and from federated AIPW in Figure 11 in Appendix B. The average treatment effects from AIPW and federated AIPW are negative and statistically significant, supporting our finding of association between the exposure to alpha blockers and a reduced risk of progression to ventilation and death.

[21] Note that these coefficient sign differences only appear in the ARD cohorts, which have fewer observations than the pneumonia cohorts.

[22] To quantify the impact of unobserved confounders, E-value analysis was conducted in Koenecke et al. (2021) which showed that a relatively strong unobserved confounder(s) would be necessary to nullify the results; see VanderWeele and Ding (2017) for the definition of E-values.

Figure 8: Federation Across MarketScan and Optum



These figures show the point estimates and confidence intervals from MarketScan and Optum as well as from our unrestricted federated MLE and unrestricted federated IPW-MLE estimators. The federated estimates of the effect of alpha blockers lie between the estimates for MarketScan and Optum for both ARD and pneumonia patients. The federated confidence intervals are in most cases narrower than those on individual databases (except for IPW-MLE on the pneumonia cohort). We use ATE weighting in IPW-MLE, and we estimate the effect of alpha blockers on reduced samples of MarketScan and Optum satisfying propensity overlap. Figure 12 in Appendix B shows the point estimates and confidence intervals with ATT weighting, which are close to those with ATE weighting. Moreover, Figure 13 in Appendix B shows the point estimates and confidence intervals from the restricted federated MLE and restricted federated IPW-MLE, which are also close to those from the unrestricted federated estimators.

6 Conclusion

This paper proposes categories of federated methods based on MLE, IPW-MLE, and AIPW that provide point estimates of treatment effects as well as variance estimates. We show that our federated point estimates have the same asymptotic distributions as the corresponding estimates from combined, individual-level data. In particular, the point estimates from our federated IPW-MLE and federated AIPW are doubly robust. We additionally show that to achieve these properties, the implementations of our federated methods should be adjusted based on conditions such as whether propensity and outcome models are correctly specified and stable across heterogeneous data sets. Finally, we suggest a general federated analysis pipeline for empirical studies and apply it to study the effectiveness of alpha blockers on patient outcomes from two separate medical claims databases.

References

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Athey, S., Chetty, R., and Imbens, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Blatt, D. and Hero, A. (2004). Distributed maximum likelihood estimation for sensor networks. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages iii–929. IEEE.

Blough, D. K., Madden, C. W., and Hornbrook, M. C. (1999). Modeling risk using generalized linear models. *Journal of health economics*, 18(2):153–171.

Blough, D. K. and Ramsey, S. D. (2000). Using generalized linear models to assess medical care costs. *Health Services and Outcomes Research Methodology*, 1(2):185–202.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188.

Du, W., Han, Y. S., and Chen, S. (2004). Privacy-preserving multivariate statistical analysis: Linear regression and classification. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 222–233. SIAM.

Fienberg, S. E., Fulp, W. J., Slavkovic, A. B., and Wrobel, T. A. (2006). "secure" log-linear and logistic regression analysis of distributed databases. In *International Conference on Privacy in Statistical Databases*, pages 277–290. Springer.

Hartung, J., Knapp, G., and Sinha, B. K. (2011). *Statistical meta-analysis with applications*, volume 738. John Wiley & Sons.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Holdcroft, A. (2007). Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine*, 100(1):2–3. PMID: 17197669.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Karr, A. F., Fulp, W. J., Vera, F., Young, S. S., Lin, X., and Reiter, J. P. (2007). Secure, privacy-preserving analysis of distributed databases. *Technometrics*, 49(3):335–345.

Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P. (2005). Secure regression on distributed databases. *Journal of Computational and Graphical Statistics*, 14(2):263–279.

Koenecke, A., Powell, M., Xiong, R., Shen, Z., Fischer, N., Huq, S., Khalafallah, A. M., Trevisan, M., Sparen, P., Carrero, J. J., et al. (2021). Alpha-1 adrenergic receptor antagonists to prevent hyperinflammation and death from lower respiratory tract infection. *Elife*, 10:e61700.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Konig, M. F., Powell, M., Staedtke, V., Bai, R.-Y., Thomas, D. L., Fischer, N., Huq, S., Khalafallah, A. M., Koenecke, A., Xiong, R., et al. (2020). Preventing cytokine storm syndrome in covid-19 using α-1 adrenergic receptor antagonists. *The Journal of clinical investigation*, 130(7):3345–3347.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60.

Li, X., Fireman, B. H., Curtis, J. R., Arterburn, D. E., Fisher, D. P., Moyneur, É., Gallagher, M., Raebel, M. A., Nowell, W. B., Lagreid, L., et al. (2019). Validity of privacy-protecting analytical methods that use only aggregate-level information to conduct multivariable-adjusted analysis in distributed data networks. *American journal of epidemiology*, 188(4):709–723.

Li, Y., Jiang, X., Wang, S., Xiong, H., and Ohno-Machado, L. (2016). Vertical grid logistic regression (vertigo). *Journal of the American Medical Informatics Association*, 23(3):570–579.

Lin, X. and Karr, A. F. (2010). Privacy-preserving maximum likelihood estimation for distributed data. *Journal of Privacy and Confidentiality*, 1(2).

Lumley, T. (2011). *Complex surveys: a guide to analysis using R*, volume 565. John Wiley & Sons.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

McMurray, R. J., Clarke, O. W., Barrasso, J. A., Clohan, D. B., Epps, Charles H., J., Glasson, J., McQuillan, R., Plows, C. W., Puzak, M. A., Orentlicher, D., and Halkola, K. A. (1991). Gender Disparities in Clinical Decision Making. *JAMA*, 266(4):559–562.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pages 2112–2245.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012.

Pollard, D. (2012). *Convergence of stochastic processes*. Springer Science & Business Media.

Powell, M., Koenecke, A., Byrd, J. B., Nishimura, A., Konig, M. F., Xiong, R., Mahmood, S., Mucaj, V., Bettegowda, C., Rose, L., et al. (2021). Ten rules for conducting retrospective pharmacoepidemiological analyses: example covid-19 study. *Frontiers in Pharmacology*, 12:1799.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

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

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121.

Rose, L., Graham, L., Koenecke, A., Powell, M., Xiong, R., Shen, Z., Mench, B., Kinzler, K. W., Bettegowda, C., Vogelstein, B., Athey, S., Vogelstein, J. T., Konig, M. F., and Wagner, T. H. (2021). The association between alpha-1 adrenergic receptor antagonists and in-hospital mortality from covid-19. *Frontiers in Medicine*, 8:304.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenman, E., Basse, G., Owen, A., and Baiocchi, M. (2020). Combining observational and experimental datasets using shrinkage estimators. *arXiv preprint arXiv:2002.06708*.

Rosenman, E., Owen, A. B., Baiocchi, M., and Banack, H. (2018). Propensity score methods for merging observational and experimental datasets. *arXiv preprint arXiv:1804.07863*.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Shu, D., Young, J. G., and Toh, S. (2019). Privacy-protecting estimation of adjusted risk ratios using modified poisson regression in multi-center studies. *BMC medical research methodology*, 19(1):1–7.

Shu, D., Young, J. G., Toh, S., and Wang, R. (2020). Variance estimation in inverse probability weighted cox models. *Biometrics*.

Singh, R. and Mukhopadhyay, K. (2011). Survival analysis in clinical trials: Basics and must know areas. *Perspectives in clinical research*, 2(4):145.

Slavkovic, A. B., Nardi, Y., and Tibbits, M. M. (2007). " secure" logistic regression of horizontally and vertically partitioned distributed databases. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 723–728. IEEE.

Snoke, J., Brick, T. R., Slavković, A., Hunter, M. D., et al. (2018). Providing accurate models across private partitioned data: Secure maximum likelihood estimation. *The Annals of Applied Statistics*, 12(2):877–914.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, 24(1):12–18.

Staedtke, V., Bai, R.-Y., Kim, K., Darvas, M., Davila, M. L., Riggins, G. J., Rothman, P. B., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., et al. (2018). Disruption of a self-amplifying catecholamine loop reduces cytokine release syndrome. *Nature*, 564(7735):273–277.

Sutton, A. J. and Higgins, J. P. (2008). Recent developments in meta-analysis. *Statistics in medicine*, 27(5):625–650.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227.

Thomsen, R. W., Christiansen, C. F., Heide-Jørgensen, U., Vogelstein, J. T., Vogelstein, B., Bettegowda, C., Tamang, S., Athey, S., and Sørensen, H. T. (2021). Association of α1-blocker receipt with 30-day mortality and risk of intensive care unit admission among adults hospitalized with influenza or pneumonia in denmark. *JAMA network open*, 4(2):e2037053–e2037053.

Toh, S., Rifas-Shiman, S. L., Lin, P.-I. D., Bailey, L. C., Forrest, C. B., Horgan, C. E., Lunsford, D., Moyneur, E., Sturtevant, J. L., Young, J. G., et al. (2020). Privacy-protecting multivariable-adjusted distributed regression analysis for multi-center pediatric study. *Pediatric research*, 87(6):1086–1092.

Toh, S., Wellman, R., Coley, R. Y., Horgan, C., Sturtevant, J., Moyneur, E., Janning, C., Pardee, R., Coleman, K. J., Arterburn, D., et al. (2018). Combining distributed regression and propensity scores: a doubly privacy-protecting analytic method for multicenter research. *Clinical Epidemiology*, 10:1773.

Tsiatis, A. A. and Davidian, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 22(4):569.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274.

Vo, T. V., Hoang, T. N., Lee, Y., and Leong, T.-Y. (2021). Federated estimation of causal effects from observational data. *arXiv preprint arXiv:2106.00456*.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in medicine*, 10(11):1665–1677.

Wolfson, M., Wallace, S. E., Masca, N., Rowe, G., Sheehan, N. A., Ferretti, V., LaFlamme, P., Tobin, M. D., Macleod, J., Little, J., et al. (2010). Datashield: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5):1372–1382.

Wooldridge, J. M. (2002). Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1(2):117–139.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281–1301.

Zhao, T. and Nehorai, A. (2007). Information-driven distributed maximum likelihood estimation based on gauss-newton method in wireless sensor networks. *IEEE Transactions on Signal Processing*, 55(9):4669–4682.

Appendices

Appendix A Additional Asymptotic Results and Practical Considerations

A.1 Additional Results

When the outcome model is unstable, if we continue using the same federation formulas as those for stable models in Section 3.1.1, Theorem 1 continues to hold for some special cases, but with $\boldsymbol{\beta}_0$ replaced by some weighted average of $\boldsymbol{\beta}_0^{(k)}$ over k.

Proposition 5 (Restricted Federated MLE for an Unstable Outcome Model)**.** *Suppose Assumptions 1.1 and 2 hold, and there exists $M < \infty$ such that $\left\|\mathcal{I}^{\text{pool}}(\boldsymbol{\beta})^{-1}\mathcal{I}^{(k)}(\boldsymbol{\beta})\right\|_2 \leq M$. Furthermore, suppose $\dot{\mathbf{d}}_y^{(k)}(\boldsymbol{\beta}) - \mathcal{I}^{(k)}(\boldsymbol{\beta}) \cdot \boldsymbol{\beta}$ and $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ do not depend on $\boldsymbol{\beta}$ for all k, where $\dot{\mathbf{d}}_y^{(k)}(\boldsymbol{\beta}) = \mathbb{E}_{(\mathbf{x},w,y)\sim\mathbb{P}^{(k)}}\left[\frac{\partial \log f(y|\mathbf{x},w;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]$. As $n_1, \cdots, n_D \to \infty$, we have*

$$n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{pool}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0^*) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d), \tag{25}$$

where $\boldsymbol{\beta}_0^*$ minimizes the Kullback-Leibler Information Criterion between $f(y|\mathbf{x}, w, \boldsymbol{\beta}_0^*)$ and the mixture of $f(y|\mathbf{x}, w, \boldsymbol{\beta}_0^{(k)})$ on the combined data. Eq. (25) *continues to hold if we replace $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}}$ by $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}}$ and/or replace $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{pool}}$ by $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}}$, where $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}}$ is estimated from the sample size pooling of the sandwich formula.*

The special cases include the linear model with i.i.d. Gaussian noise that has variance σ_e^2. In this case, $\mathcal{I}^{(k)}(\boldsymbol{\beta}) = \mathbf{X}^\top\mathbf{X}/\sigma_e^2$ and $\dot{\mathbf{d}}_y^{(k)}(\boldsymbol{\beta}) - \mathcal{I}^{(k)}(\boldsymbol{\beta}) \cdot \boldsymbol{\beta} = -\mathbf{Y}^\top\mathbf{X}/\sigma_e^2$ do not depend on $\boldsymbol{\beta}$; therefore, the assumptions in Proposition 5 are satisfied. Moreover, $\boldsymbol{\beta}_0^*$ is a linear combination of $(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\beta}_0^{(2)}, \cdots, \boldsymbol{\beta}_0^{(D)})$ that satisfies $\sum_{k=1}^D p_k \mathbb{E}_{\mathbf{x}\sim\mathbb{P}^{(k)}}[\mathbf{x}] \cdot (\boldsymbol{\beta}_0^{(k)} - \boldsymbol{\beta}_0^*) = 0$. Hence, the pooled model parameters are a weighted average of the estimated parameters on individual data sets.

A.2 Practical Considerations

Remark 3. Suppose (1) Y_i is a binary response variable (such as the indicator variable for ventilation and death in our empirical medical example); (2) $\mathbb{E}[Y_i|\mathbf{X}_i]$ follows a logistic regression model; (3) we use the true propensity score in the IPW-MLE. Then the estimated $\mathbf{D}_{\boldsymbol{\beta}_0,\varpi}^{(k)}$ on data set k, denoted as $\hat{\mathbf{D}}_{\boldsymbol{\beta},\varpi}^{(k)}$, equals

$$\hat{\mathbf{D}}_{\boldsymbol{\beta},\varpi}^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k}\Big(\frac{W_i}{(\hat{e}_i^{\text{fed}})^2} + \frac{1-W_i}{(1-\hat{e}_i^{\text{fed}})^2}\Big)\hat{\varepsilon}_i^2\mathbf{X}_i\mathbf{X}_i^\top,$$

where ε_i is unit i's residual. Some commonly used packages, such as `syvglm` in `R` (Lumley, 2011), use working residuals for $\hat{\varepsilon}_i$ (i.e., $\hat{\varepsilon}_i = \frac{Y_i - \hat{p}_i}{\hat{p}_i(1-\hat{p}_i)}$ and $\hat{p}_i = \frac{\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}})}{1+\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}})}$).

Appendix B Additional Empirical Results

B.1 A Toy Example for Inverse Variance Weighting

In this section, we present a simplified example for the federated treatment coefficient from inverse variance weighting lying outside the interval between treatment coefficients on two data sets. Suppose we only have treatment and age in the outcome model, and the coefficients and inverse variance matrices on two data sets [23] are:

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = \begin{bmatrix} \hat{\beta}_{\mathrm{M},w} \\ \hat{\beta}_{\mathrm{M,age}} \end{bmatrix} = \begin{bmatrix} -0.67 \\ 2.03 \end{bmatrix}, \quad \hat{\mathbf{V}}_{\mathrm{M}}^{-1} = \begin{bmatrix} 51.6 & -28.6 \\ -28.6 & 474.02 \end{bmatrix},$$
$$\hat{\boldsymbol{\beta}}_{\mathrm{O}} = \begin{bmatrix} \hat{\beta}_{\mathrm{O},w} \\ \hat{\beta}_{\mathrm{O,age}} \end{bmatrix} = \begin{bmatrix} -0.02 \\ -0.15 \end{bmatrix}, \quad \hat{\mathbf{V}}_{\mathrm{O}}^{-1} = \begin{bmatrix} 55.34 & 14.61 \\ 14.61 & 187.98 \end{bmatrix}.$$

Then the federated coefficients based on inverse variance weighting are

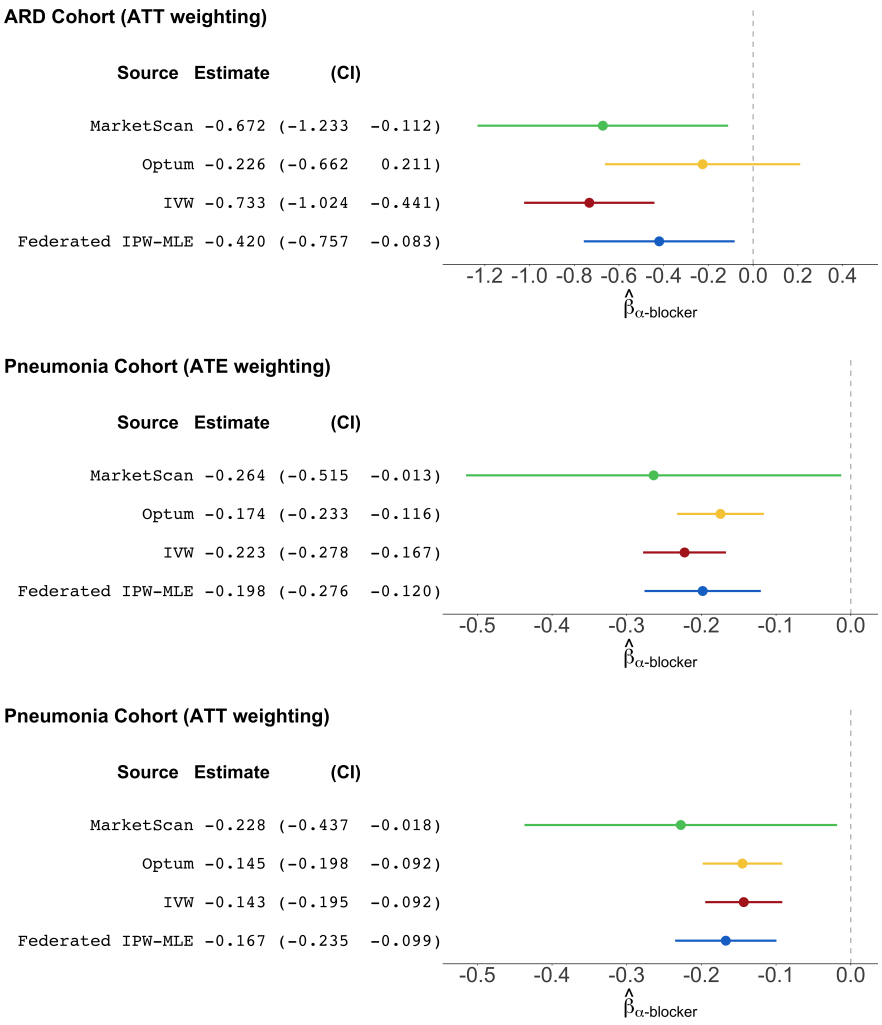
$$\hat{\boldsymbol{\beta}}_{\mathrm{ivw}} = \left(\hat{\mathbf{V}}_{\mathrm{M}}^{-1} + \hat{\mathbf{V}}_{\mathrm{O}}^{-1}\right)^{-1}\left(\hat{\mathbf{V}}_{\mathrm{M}}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{M}} + \hat{\mathbf{V}}_{\mathrm{O}}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{O}}\right) = \begin{bmatrix} -0.71 \\ 1.42 \end{bmatrix}$$

The federated treatment coefficient is -0.71, which is smaller than $\hat{\beta}_{\mathrm{M},w}$ and $\hat{\beta}_{\mathrm{O},w}$.

B.2 Additional Results for Federation Across Two Medical Claim Data Sets

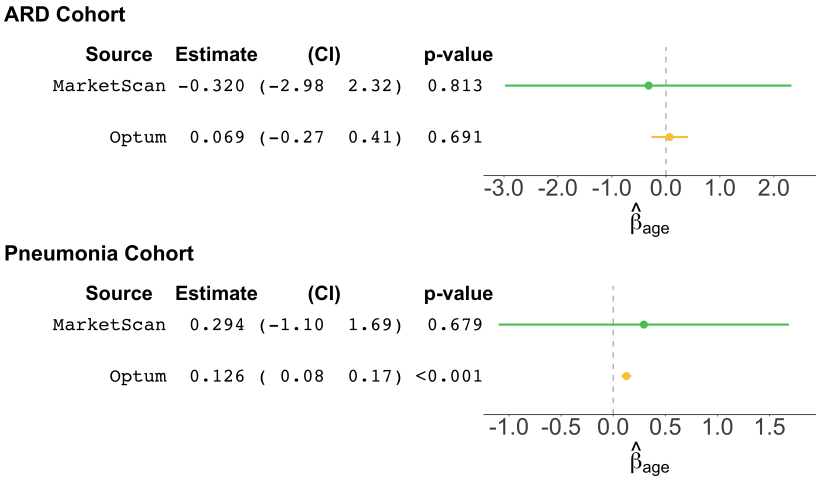
[23] These numbers are identical to those in the inverse propensity-weighted logistic regression on MarketScan and Optum ARD cohorts.

Figure 9: Coefficient of the Exposure to Alpha Blockers

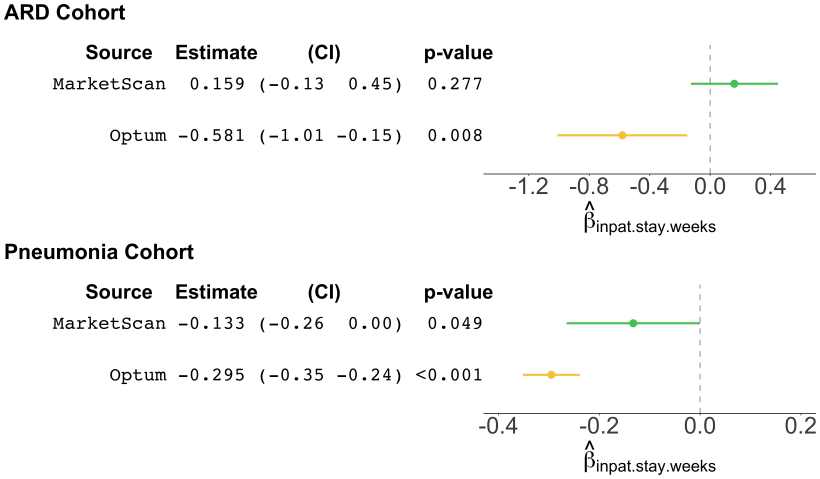


We plot the coefficient of the exposure to alpha blockers in an inverse propensity-weighted logistic regression to model the probability of progression to mechanical ventilation and then death of inpatients with acute respiratory distress (or pneumonia). The top two lines in each panel show the estimated coefficient and 95% confidence interval on MarketScan and Optum, respectively. The third and fourth lines correspond to the federated coefficient and 95% confidence interval of MarketScan and Optum from IVW, and from our unrestricted federated IPW-MLE, respectively. While the federated coefficient from our unrestricted federated IPW-MLE always lies between those from MarketScan and Optum, the federated coefficient from IVW does not. We consider both ATE and ATT weighting in IPW-MLE (see Figure 1 for the ATE weighting in IPW-MLE for the ARD cohort), and we estimate the effect of alpha blockers on reduced samples of MarketScan and Optum satisfying propensity overlap.

Figure 10: Estimated Coefficients and Confidence Intervals for Selected Confounders in the Logistic Regression Model

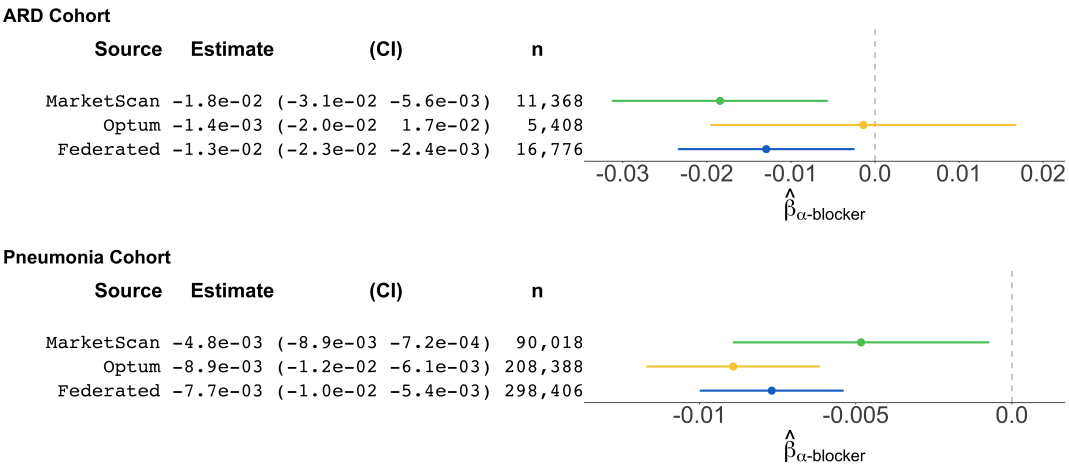


(a) Coefficient for age

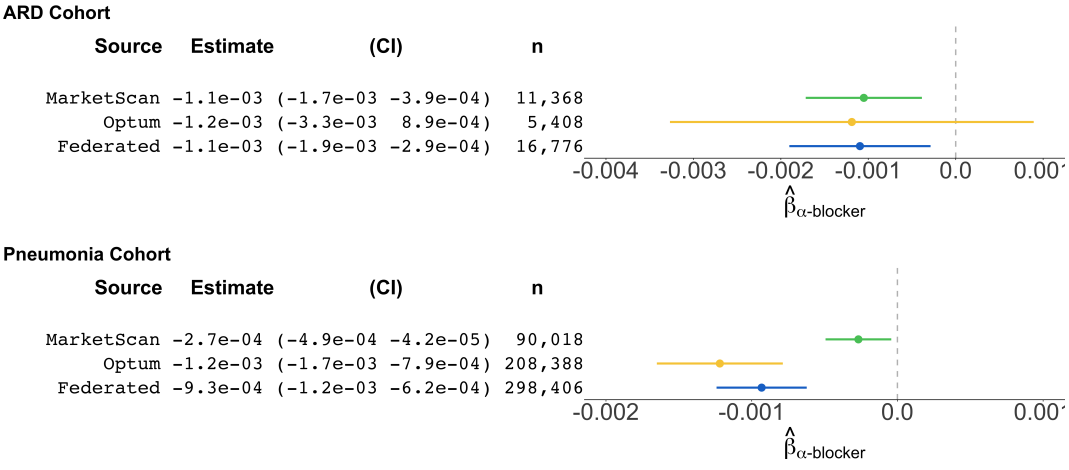


(b) Coefficient for weeks with prior inpatient admissions

Figure 11: Federation of AIPW Across MarketScan and Optum



(a) ATE



(b) ATT

This figure shows the estimated ATE and ATT, and their estimated confidence intervals from AIPW on MarketScan and Optum as well as from our unrestricted federated AIPW estimator. The federated estimates of the effect of alpha blockers lie between the estimates on MarketScan and Optum for both ARD and pneumonia patients. The federated confidence intervals are narrower than those on individual databases on both ARD and pneumonia cohorts.

Table 6: Comparison Between Restricted/Unrestricted Federated IPW-MLE and IVW with Corresponding Restricted/Unrestricted Benchmarks and ATT Weighting

(a) IPW-MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

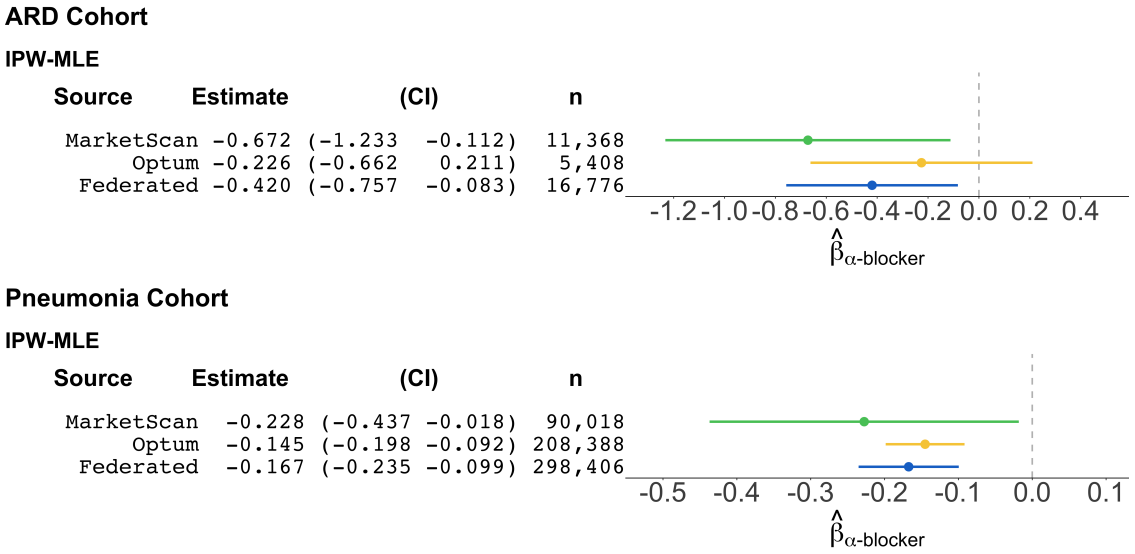
	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
MS ARD	*-0.7324*	0.7594	0.0584	0.0597
MS PNA	*-0.2390*	0.2488	0.0223	0.0229
Optum PNA	*-0.1524*	0.0263	0.0061	0.0068
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
MS ARD	*0.0812*	0.0508	0.0206	0.0077
MS PNA	*0.0174*	0.0031	0.0009	0.0006
Optum PNA	*0.0026*	0.0001	0.0000	0.0000

(b) IPW-MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
MS ARD	*-0.7659*	0.7260	0.0888	0.0511
MS PNA	*-0.2387*	0.2492	0.0225	0.0225
Optum PNA	*-0.1528*	0.0260	0.0059	0.0062
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
MS ARD	*0.0769*	0.0465	0.0163	0.0053
MS PNA	*0.0174*	0.0032	0.0009	0.0006
Optum PNA	*0.0026*	0.0001	0.0000	0.0000

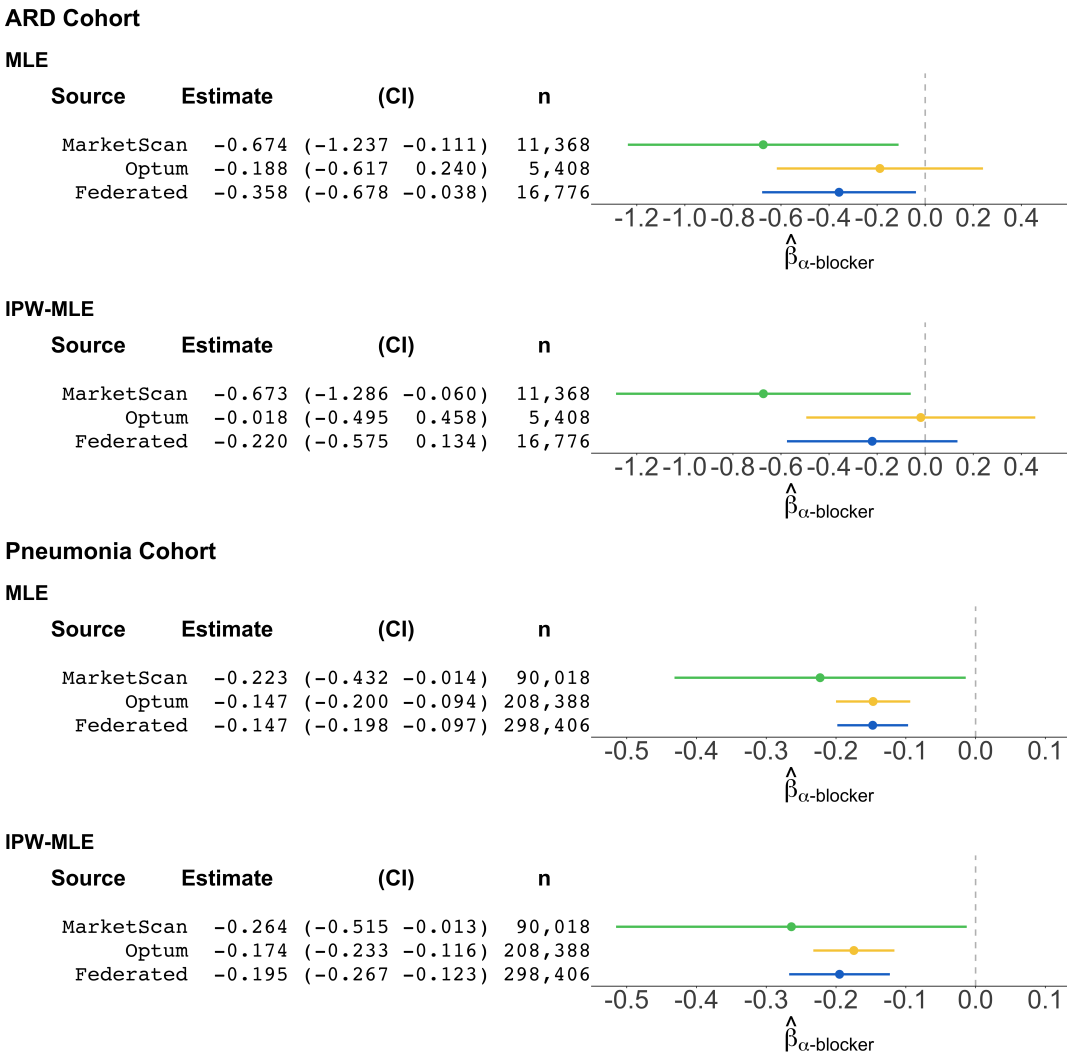
These tables compare restricted and unrestricted federated IPW-MLE (restricted: $\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ and $\hat{V}^{\mathbf{r}.\text{fed}}_{\beta_w,\text{ipw-mle}}$; $\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ and $\hat{V}^{\mathbf{unr}.\text{fed}}_{\beta_w,\text{ipw-mle}}$) estimators, and inverse variance weighted ($\hat{\beta}_{w,\text{ivw}}$ and $\hat{V}_{\beta_w,\text{ivw}}$) estimators to the restricted and unrestricted benchmarks (restricted: $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{r}}_{\beta_w,\text{bm}}$; unrestricted: $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{unr}}_{\beta_w,\text{bm}}$). In the federated IPW-MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\gamma}}$ which allows for propensity model misspecification (i.e., when Condition 7 is violated). We construct subsamples from the MarketScan ARD (MS ARD) cohort, and from the MarketScan and Optum pneumonia (MS PNA and Optum PNA) cohorts, letting $D = 2$. For MS ARD, $n_1 = n_2 = n_{\text{pool}}/2 = 6,000$; for MS PNA and Optum PNA, $n_1 = n_2 = n_{\text{pool}}/2 = 30,000$. We use ATT weighting in IPW-MLE. The mean absolute error (MAE) is calculated relative to the benchmark mean values (first column of each table) based on 50 iterations of independent sampling of subsamples. Note that the combined data $\mathcal{C}_1 \cup \mathcal{C}_2$ vary across iterations, so we report the average of the benchmarks ($\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{bm}}$ and $\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{bm}}$) across iterations, yielding slight noise in the benchmark means reported.

Figure 12: Federation Across MarketScan and Optum (Unrestricted Federated Estimators with ATT Weighting)



These figures show the point estimates and confidence intervals from MarketScan and Optum as well as from our **unrestricted** federated IPW-MLE estimators. The federated estimates of the effect of alpha blockers lie between the estimates for MarketScan and Optum for both ARD and pneumonia patients. We use **ATT weighting** in IPW-MLE, and we estimate the effect of alpha blockers on reduced samples of MarketScan and Optum satisfying propensity overlap. The results are similar to those with ATE weighting in Figure 8.

Figure 13: Federation Across MarketScan and Optum (Restricted Federated Estimators with ATE Weighting)



These figures show the point estimates and confidence intervals from MarketScan and Optum as well as from our **restricted** federated MLE and **restricted** federated IPW-MLE estimators. The federated estimates of the effect of alpha blockers lie between the estimates for MarketScan and Optum for both ARD and pneumonia patients. The federated confidence intervals are generally narrower than those on individual databases (except for IPW-MLE on the pneumonia cohort). We use **ATE weighting** in IPW-MLE, and we estimate the effect of alpha blockers on reduced samples of MarketScan and Optum satisfying propensity overlap.

B.3 Federated Results Based on Alternative Sampling Schemes from One Medical Claims Data Set

In this section, we consider alternative approaches to construct subsamples. The results are presented in Tables 7-10, and are consistent with the results in Section 5.2.

Varying Sampling Ratios of Sub-cohorts We consider the cases where 50% (70% or 90%) of the patient records in $\mathcal{S}_1$ are sampled from $\mathcal{C}_1$ with replacement, with the remaining 50% (30% or 10%) sampled from $\mathcal{C}_2$ with replacement. Similarly, 50% (30% or 10%) of the patient records in $\mathcal{S}_2$ are sampled from $\mathcal{C}_1$ with replacement, with the remaining 50% (70% or 90%) sampled from $\mathcal{C}_2$ with replacement. Therefore, for the 50%/50% sampling scheme, the age structure in $\mathcal{S}_1$ and $\mathcal{S}_2$ are similar; for other sampling schemes, $\mathcal{S}_1$ has more young patients than $\mathcal{S}_2$. The results are presented in Table 7.

Sub-cohorts Based on Patient Age and Fiscal Year We construct two equally-sized subsamples. We first split the cohort $\mathcal{C}$ into four sub-cohorts, denoted as $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, and $\mathcal{C}_4$, by age and fiscal year ($\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cup \mathcal{C}_4$ and $\mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3 \cap \mathcal{C}_4 = \emptyset$). $\mathcal{C}_1$ only has records of the patients whose age is below the median of all patients in $\mathcal{C}$, and fiscal year is from 2004 through 2012; $\mathcal{C}_2$ only has records of the patients whose age is below the median of all patients in $\mathcal{C}$, and fiscal year is from 2013 through 2016; $\mathcal{C}_3$ only has records of the patients whose age is above the median of all patients in $\mathcal{C}$, and fiscal year is from 2004 through 2012; $\mathcal{C}_4$ only has records of the patients whose age is above the median of all patients in $\mathcal{C}$, and fiscal year is from 2013 through 2016. The two subsamples are constructed as follows: 70% of the patient records in $\mathcal{S}_1$ are sampled from $\mathcal{C}_1$ with replacement, and 10% are sampled from $\mathcal{C}_k$ with replacement for all $k \neq 1$; Similarly, 70% of the patient records in $\mathcal{S}_2$ are sampled from $\mathcal{C}_4$ with replacement, and 10% are sampled from $\mathcal{C}_k$ with replacement for all $k \neq 4$. The results are presented in Table 8.

Varying Subsample Sizes We follow the same sampling schemes as Section 5.2 and construct two subsamples with different sizes (i.e. $(n_1, n_2) \in \{(40,000, 10,000), (20,000, 10,000)\}$). The results are presented in Table 9.

Varying Number of Subsamples We construct D equally-sized subsamples, where D varies from 2 to 4. Similar as Section 5.2, we split the cohort $\mathcal{C}$ into D sub-cohorts, where $\mathcal{C}_j$ has records of the patients who age is between $(j-1)/D$ and j/D percentiles for all j. We construct subsamples $\mathcal{S}_j$ as follows: 70% of the patient records in $\mathcal{S}_j$ are sampled from $\mathcal{C}_j$ with replacement, and 30/(D-1)% are sampled from $\mathcal{C}_k$ with replacement for all $k \neq j$. The results are presented in Table 10.

Table 7: Comparison Between Restricted/Unrestricted Federated Estimators and IVW with the Corresponding Restricted/Unrestricted Benchmarks with Varying Sampling Ratios of Sub-cohorts

(a) MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
50%/50%	*-0.2077*	0.0504	0.0246	0.0287
70%/30%	*-0.1883*	0.0586	0.0289	0.0305
90%/10%	*-0.2294*	0.0503	0.0262	0.0301
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
50%/50%	*0.0515*	0.0081	0.0043	0.0022
70%/30%	*0.0508*	0.0086	0.0047	0.0024
90%/10%	*0.0528*	0.0082	0.0049	0.0022

(b) MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
50%/50%	*-0.2071*	0.0517	0.0272	0.0270
70%/30%	*-0.1882*	0.0582	0.0318	0.0296
90%/10%	*-0.2276*	0.0498	0.0274	0.0268
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
50%/50%	*0.0517*	0.0083	0.0045	0.0023
70%/30%	*0.0509*	0.0088	0.0049	0.0025
90%/10%	*0.0530*	0.0083	0.0051	0.0023

(c) IPW-MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
50%/50%	*-0.2793*	0.7466	0.0322	0.0299
70%/30%	*-0.2630*	0.7721	0.0383	0.0383
90%/10%	*-0.3029*	0.8316	0.0342	0.0471
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
50%/50%	*0.0697*	0.0449	0.0167	0.0063
70%/30%	*0.0705*	0.0467	0.0176	0.0074
90%/10%	*0.0720*	0.0492	0.0177	0.0077

(d) IPW-MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
50%/50%	*-0.2777*	0.7482	0.0391	0.0088
70%/30%	*-0.2668*	0.7683	0.0525	0.0118
90%/10%	*-0.3121*	0.8224	0.0518	0.0137
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
50%/50%	*0.0666*	0.0418	0.0136	0.0020
70%/30%	*0.0666*	0.0428	0.0138	0.0023
90%/10%	*0.0675*	0.0448	0.0133	0.0027

These tables compare restricted and unrestricted federated MLE and federated IPW-MLE, and IVW with the restricted and unrestricted benchmarks, where the sub-cohorts are constructed with varying sampling ratios of sub-cohorts. In the federated MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\beta}}$ that is appropriate when the outcome model is misspecified. In the federated IPW-MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\gamma}}$ that allows for propensity model misspecification (where Condition (7) is violated). We construct subsamples from the MarketScan pneumonia (MS PNA) cohort. We let $D = 2$ with $n_1 = n_2 = n_{\text{pool}}/2 = 10,000$. We use ATE weighting in IPW-MLE. The benchmark means and MAE are calculated based on 50 iterations.

Table 8: Comparison Between Restricted/Unrestricted Federated Estimators and IVW with the Corresponding Restricted/Unrestricted Benchmarks with Sub-cohorts Based on Patient Age and Fiscal Year

(a) MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\mathrm{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\mathrm{bm}}$

$\hat{\beta}^{\mathbf{r}}_{w,\mathrm{bm}}$ **mean**	$\hat{\beta}_{w,\mathrm{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\mathrm{fed}}_{w,\mathrm{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\mathrm{fed}}_{w,\mathrm{mle}}$ **MAE**
-0.1876	0.0281	0.0163	0.0182
$\hat{V}^{\mathbf{r}}_{w,\mathrm{bm}}$ **mean**	$\hat{V}_{w,\mathrm{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\mathrm{fed}}_{w,\mathrm{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\mathrm{fed}}_{w,\mathrm{mle}}$ **MAE**
0.0227	0.0023	0.0012	0.0006

(b) MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\mathrm{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\mathrm{bm}}$

$\hat{\beta}^{\mathbf{unr}}_{w,\mathrm{bm}}$ **mean**	$\hat{\beta}_{w,\mathrm{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\mathrm{fed}}_{w,\mathrm{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\mathrm{fed}}_{w,\mathrm{mle}}$ **MAE**
-0.1858	0.0285	0.0163	0.0156
$\hat{V}^{\mathbf{unr}}_{w,\mathrm{bm}}$ **mean**	$\hat{V}_{w,\mathrm{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\mathrm{fed}}_{w,\mathrm{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\mathrm{fed}}_{w,\mathrm{mle}}$ **MAE**
0.0228	0.0023	0.0012	0.0006

(c) IPW-MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\mathrm{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\mathrm{bm}}$

$\hat{\beta}^{\mathbf{r}}_{w,\mathrm{bm}}$ **mean**	$\hat{\beta}_{w,\mathrm{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\mathrm{fed}}_{w,\mathrm{ipw\text{-}mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\mathrm{fed}}_{w,\mathrm{ipw\text{-}mle}}$ **MAE**
-0.2662	0.5273	0.0155	0.0162
$\hat{V}^{\mathbf{r}}_{w,\mathrm{bm}}$ **mean**	$\hat{V}_{w,\mathrm{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\mathrm{fed}}_{w,\mathrm{ipw\text{-}mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\mathrm{fed}}_{w,\mathrm{ipw\text{-}mle}}$ **MAE**
0.0348	0.0182	0.0049	0.0020

(d) IPW-MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\mathrm{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\mathrm{bm}}$

$\hat{\beta}^{\mathbf{unr}}_{w,\mathrm{bm}}$ **mean**	$\hat{\beta}_{w,\mathrm{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\mathrm{fed}}_{w,\mathrm{ipw\text{-}mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\mathrm{fed}}_{w,\mathrm{ipw\text{-}mle}}$ **MAE**
-0.2686	0.5249	0.0243	0.0073
$\hat{V}^{\mathbf{unr}}_{w,\mathrm{bm}}$ **mean**	$\hat{V}_{w,\mathrm{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\mathrm{fed}}_{w,\mathrm{ipw\text{-}mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\mathrm{fed}}_{w,\mathrm{ipw\text{-}mle}}$ **MAE**
0.0340	0.0174	0.0041	0.0009

These tables compare restricted and unrestricted federated MLE and federated IPW-MLE, and IVW with the restricted and unrestricted benchmarks, where the sub-cohorts are constructed based on patient age and fiscal year. In the federated MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\beta}}$ that is appropriate when the outcome model is misspecified. In the federated IPW-MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\gamma}}$ that allows for propensity model misspecification (where Condition (7) is violated). We construct subsamples from the MarketScan pneumonia (MS PNA) cohort. We let $D = 2$ with $n_1 = n_2 = n_{\mathrm{pool}}/4 = 20,000$. We use ATE weighting in IPW-MLE. The benchmark means and MAE are calculated based on 50 iterations.

Table 9: Comparison Between Restricted/Unrestricted Federated Estimators and IVW with the Corresponding Restricted/Unrestricted Benchmarks with Varying Subsample Sizes

(a) MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
20k10k	*-0.2379*	0.0348	0.0162	0.0164
40k10k	*-0.2688*	0.0259	0.0102	0.0106
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
20k10k	*0.0372*	0.0043	0.0025	0.0009
40k10k	*0.0243*	0.0018	0.0011	0.0004

(b) MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
20k10k	*-0.2369*	0.0342	0.0173	0.0137
40k10k	*-0.2681*	0.0254	0.0117	0.0089
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
20k10k	*0.0372*	0.0043	0.0026	0.0009
40k10k	*0.0243*	0.0019	0.0011	0.0004

(c) IPW-MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
20k10k	*-0.3444*	0.6105	0.0527	0.0581
40k10k	*-0.3590*	0.4971	0.0898	0.0912
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
20k10k	*0.0508*	0.0291	0.0105	0.0050
40k10k	*0.0342*	0.0175	0.0048	0.0048

(d) IPW-MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
20k10k	*-0.3507*	0.6042	0.0620	0.0576
40k10k	*-0.3652*	0.4910	0.0974	0.0961
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
20k10k	*0.0497*	0.0279	0.0093	0.0044
40k10k	*0.0341*	0.0174	0.0052	0.0052

These tables compare restricted and unrestricted federated MLE and federated IPW-MLE, and IVW with the restricted and unrestricted benchmarks with varying subsample sizes. In the federated MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\beta}}$ that is appropriate when the outcome model is misspecified. In the federated IPW-MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\gamma}}$ that allows for propensity model misspecification (where Condition (7) is violated). We construct subsamples from the MarketScan pneumonia (MS PNA) cohort. We let $D = 2$ with varying values of n_1 and n_2. We use ATE weighting in IPW-MLE. The benchmark means and MAE are calculated based on 50 iterations.

Table 10: Comparison Between Restricted/Unrestricted Federated Estimators and IVW with the Corresponding Restricted/Unrestricted Benchmarks with Varying Number of Subsamples

(a) MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
$D=2$	*-0.2100*	0.0312	0.0168	0.0179
$D=3$	*-0.2096*	0.0303	0.0146	0.0171
$D=4$	*-0.2345*	0.0354	0.0255	0.0277
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
$D=2$	*0.0342*	0.0039	0.0020	0.0009
$D=3$	*0.0228*	0.0031	0.0016	0.0005
$D=4$	*0.0172*	0.0031	0.0017	0.0007

(b) MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
$D=2$	*-0.2098*	0.0323	0.0187	0.0161
$D=3$	*-0.2080*	0.0315	0.0147	0.0151
$D=4$	*-0.2328*	0.0366	0.0247	0.0259
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
$D=2$	*0.0342*	0.0040	0.0021	0.0010
$D=3$	*0.0228*	0.0032	0.0016	0.0006
$D=4$	*0.0173*	0.0031	0.0017	0.0008

(c) IPW-MLE: Restricted Benchmarks $\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{r}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
$D=2$	*-0.2757*	0.5992	0.0199	0.0251
$D=3$	*-0.2461*	0.8752	0.0230	0.0282
$D=4$	*-0.2671*	0.9915	0.0335	0.0280
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
$D=2$	*0.0471*	0.0264	0.0082	0.0032
$D=3$	*0.0327*	0.0229	0.0079	0.0026
$D=4$	*0.0241*	0.0180	0.0073	0.0022

(d) IPW-MLE: Unrestricted Benchmarks $\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$, $\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
$D=2$	*-0.2815*	0.5933	0.0328	0.0045
$D=3$	*-0.2566*	0.8647	0.0364	0.0071
$D=4$	*-0.2843*	0.9743	0.0525	0.0088
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
$D=2$	*0.0452*	0.0246	0.0063	0.0008
$D=3$	*0.0310*	0.0212	0.0062	0.0006
$D=4$	*0.0229*	0.0169	0.0056	0.0006

These tables compare restricted and unrestricted federated MLE and federated IPW-MLE, and IVW with the restricted and unrestricted benchmarks with varying number of subsamples. In the federated MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\beta}}$ that is appropriate when the outcome model is misspecified. In the federated IPW-MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\gamma}}$ that allows for propensity model misspecification (where Condition (7) is violated). We construct subsamples from the MarketScan pneumonia (MS PNA) cohort. We let $D \in \{2,3,4\}$ with $n_j = 15,000$ for all $j \in \{1,\cdots,D\}$ (Note that $D=2$ and $n_j = 30,000$ in Table 5, and therefore the MAE values in this table are larger than those in Table 5). We use ATE weighting in IPW-MLE. The benchmark means and MAE are calculated based on 50 iterations.

Appendix C Data-Driven Simulations based on Medical Claims Data

In this section, we conduct semi-synthetic retrospective analyses based on the cohorts of patients described in Section 5. Similar to the analyses in Section 5, we sample patient records to construct subsamples and compare various federated estimators. But, in contrast to the analyses in Section 5, only the covariate information in patient records is kept in subsamples, and patients' treatment assignment and outcome are simulated based on their covariates using known propensity and outcome models. In this way, we can control for whether the conditions in Section 2.3 hold or not.

In Section C.1, we present various sampling schemes to simulate various conditions; results are presented in Section C.2.

C.1 Sampling Schemes for Subsamples

We construct D subsamples, with n_j patient records in subsample j, based on the patients' covariates from cohort $\mathcal{C}$ (here, we focus on $\mathcal{C}_{\text{M,ARD}}$). Let $\mathcal{C}_j \subseteq \mathcal{C}$ be a sub-cohort of $\mathcal{C}$; we construct subsample j based on $\mathcal{C}_j$. Let $\mathcal{C}_{j,\mathbf{x}} = \{\mathbf{x} : (\mathbf{x}, w, y) \in \mathcal{C}_j\}$ be the set of patient covariates. We construct subsample j as follows. First, we sample one patient's covariates from $\mathcal{C}_{j,\mathbf{x}}$. Second, we sample this patient's binary treatment status W from a Bernoulli distribution with treated probability $\text{pr}(W = 1|\mathbf{X})$ calculated from subsample j's propensity model $e^{(j)}(\mathbf{X})$. Third, we sample this patient's binary outcome Y from a Bernoulli distribution with outcome probability $\text{pr}(Y = 1|\mathbf{X}, W)$ calculated from subsample j's outcome model $f^{(j)}(Y|\mathbf{X}, W)$. For our simulations, we repeat the above three steps n_j times, sampling patient covariates with replacement in the first step.

Stability of Propensity and Outcome Models Let $e^{(j)}(\mathbf{X})$ and $f^{(j)}(Y|\mathbf{X}, W)$ denote our estimated propensity model and outcome model, respectively, on cohort $\mathcal{C}_j$. We then describe the two cases for simulated model stability. The first case is for stable propensity and outcome models, where we let $\mathcal{C}_j = \mathcal{C}_k = \mathcal{C}$ for all $j \neq k$. The second case is for unstable propensity and outcome models, where we partition $\mathcal{C}$ into $D = 2$ sub-cohorts based on the patient age covariate. In the results presented in Section C.2, we generate $\mathcal{C}_1$ such that it is comprised of 95% records with patient ages below the $\mathcal{C}$ median of patient ages, and 5% records with patient ages above the median. $\mathcal{C}_2$ is the complement of $\mathcal{C}_1$.

Propensity and Outcome Model Specifications We consider two model specifications for the true propensity model $e^{(j)}(\mathbf{X})$: one is the logistic regression model and the other is the classification tree model. Similarly, we consider two model specifications for the true outcome model $f^{(j)}(Y|\mathbf{X}, W)$: one is the logistic regression model and the other is the classification tree model. In our federated MLE and IPW-MLE analyses, we estimate both the propensity and outcome model from logistic regressions. Therefore, we correctly specify the propensity (or outcome) model

if $e^{(j)}(\mathbf{X})$ (or $f^{(j)}(Y|\mathbf{X}, W)$) follows a logistic regression model, and we misspecify the propensity/outcome model if $e^{(j)}(\mathbf{X})$ (or $f^{(j)}(Y|\mathbf{X}, W)$) follows a classification tree model.

C.2 Results

In this section, we compare our restricted and unrestricted federated estimators (MLE and PW-MLE) and IVW with the benchmarks. The results in this section confirm our three main findings in Section 5.2.3.

Table 11: Data-Driven Simulations: Comparison Between Restricted/Unrestricted Federated Estimators and IVW with the Corresponding Restricted/Unrestricted Benchmarks

(a) MLE: Stable propensity and outcome models

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**
correct spec.	*-0.6748*	0.0190	0.0114
misspec. outcome	*-0.0552*	0.0130	0.0060
misspec. propensity	*-0.6750*	0.0187	0.0117
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$ **MAE**
correct spec.	*0.0237*	0.0016	0.0008
misspec. outcome	*0.0139*	0.0007	0.0003
misspec. propensity	*0.0232*	0.0016	0.0008

(b) MLE: Unstable propensity and outcome models

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
correct spec.	*-0.7516*	0.0832	0.0417
misspec. outcome	*-0.0176*	0.0140	0.0056
misspec. propensity	*-0.7288*	0.0905	0.0353
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$ **MAE**
correct spec.	*0.0251*	0.0077	0.0018
misspec. outcome	*0.0127*	0.0007	0.0001
misspec. propensity	*0.0236*	0.0067	0.0014

(c) IPW-MLE: Stable propensity and outcome models

	$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
correct spec.	*-0.6877*	0.5044	0.0172
misspec. outcome	*-0.0642*	0.3433	0.0082
misspec. propensity	*-0.6673*	0.4514	0.0163
	$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
correct spec.	*0.0356*	0.0192	0.0039
misspec. outcome	*0.0216*	0.0083	0.0022
misspec. propensity	*0.0296*	0.0133	0.0028

(d) IPW-MLE: Unstable propensity and outcome models

	$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{\beta}_{w,\text{ivw}}$ **MAE**	$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
correct spec.	*-0.7679*	0.5166	0.0236
misspec. outcome	*-0.0249*	0.2962	0.0008
misspec. propensity	*-0.7648*	0.4177	0.0188
	$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$ **mean**	$\hat{V}_{w,\text{ivw}}$ **MAE**	$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$ **MAE**
correct spec.	*0.0346*	0.0176	0.0017
misspec. outcome	*0.0189*	0.0062	0.0001
misspec. propensity	*0.0279*	0.0111	0.0011

These tables compare restricted (or unrestricted) federated MLE and federated IPW-MLE, and IVW with the restricted (or unrestricted) benchmarks for stable and unstable propensity and outcome models, and for correctly specified propensity and outcome models (i.e., "correct spec."), correctly specified propensity and misspecified outcome models (i.e., "misspec. outcome"), misspecified propensity and correctly specified outcome models (i.e., "misspec. propensity"). For the federated MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\beta}}$. For the federated IPW-MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\gamma}}$ that allows for propensity model misspecification (where Condition (7) is violated). We construct subsamples from the MarketScan ARD cohort. We let $D = 2$ with $n_1 = n_2 = 20,000$. We use ATE weighting for IPW-MLE.

Table 12: Data-Driven Simulations: Comparison Between Restricted/Unrestricted Federated Estimators with the Corresponding Restricted/Unrestricted Benchmarks

(a) MLE: Stable propensity and outcome models

$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$		$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$	
mean	**mean**	**MAE**	**mean**	**MAE**
-0.6748	-0.6635	0.0114	-0.6644	0.0116
$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$		$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$	
mean	**mean**	**MAE**	**mean**	**MAE**
0.0237	0.0230	0.0008	0.0233	0.0005

(b) MLE: Unstable propensity and outcome models

$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$		$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$	
mean	**mean**	**MAE**	**mean**	**MAE**
-0.7516	-0.6762	0.0758	-0.7099	0.0417
$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{mle}}$		$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{mle}}$	
mean	**mean**	**MAE**	**mean**	**MAE**
0.0251	0.0203	0.0048	0.0233	0.0018

(c) IPW-MLE: Stable propensity and outcome models

$\hat{\beta}^{\mathbf{r}}_{w,\text{bm}}$	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$		$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$	
mean	**mean**	**MAE**	**mean**	**MAE**
-0.6877	-0.671	0.0172	-0.6849	0.0190
$\hat{V}^{\mathbf{r}}_{w,\text{bm}}$	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$		$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$	
mean	**mean**	**MAE**	**mean**	**MAE**
0.0356	0.0316	0.0039	0.0335	0.0024

(d) IPW-MLE: Unstable propensity and outcome models

$\hat{\beta}^{\mathbf{unr}}_{w,\text{bm}}$	$\hat{\beta}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$		$\hat{\beta}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$	
mean	**mean**	**MAE**	**mean**	**MAE**
-0.7679	-0.6879	0.0862	-0.7443	0.0236
$\hat{V}^{\mathbf{unr}}_{w,\text{bm}}$	$\hat{V}^{\mathbf{r}.\text{fed}}_{w,\text{ipw-mle}}$		$\hat{V}^{\mathbf{unr}.\text{fed}}_{w,\text{ipw-mle}}$	
mean	**mean**	**MAE**	**mean**	**MAE**
0.0346	0.0277	0.0070	0.0329	0.0017

These tables compare restricted and unrestricted federated MLE and federated IPW-MLE with the restricted (or unrestricted) benchmarks for stable and unstable propensity and outcome models. The propensity model is correctly specified in these tables. For the federated MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\beta}}$. For the federated IPW-MLE, we use the sandwich formula for $\mathbf{V}_{\boldsymbol{\gamma}}$. We construct subsamples from the MarketScan ARD cohort. We let $D = 2$ with $n_1 = n_2 = 20,000$. We use ATE weighting for IPW-MLE.

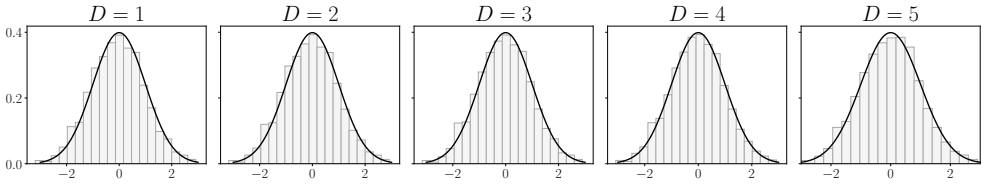
Appendix D Simulations for Finite-Sample Properties

In this section, we demonstrate the finite sample properties of our asymptotic results for the federated MLE, federated IPW-MLE, and federated AIPW, and confirm our theoretical distribution results. To conserve space, we present the finite-sample results for the case in which the propensity and outcome models are stable, estimated, and correctly specified (where Conditions 3-8 hold). The results for other cases are similar and available upon request. In our data generating process, $\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \text{unif}(-1, 1)$ is a scalar, and Y_i is a binary response variable that follows

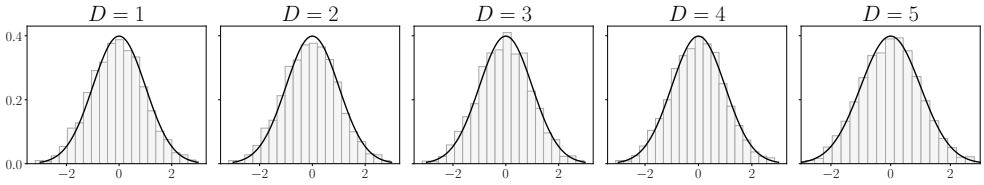
$$\frac{p_y}{1-p_y} = \exp(\beta_c + \beta_w W_i + \beta_x \mathbf{X}_i), \qquad p_y = P(Y_i = 1)$$
$$\frac{p_e}{1-p_e} = \exp(\gamma_c + \gamma_x \mathbf{X}_i), \qquad p_e = \text{pr}(W_i = 1)$$

where $\boldsymbol{\beta}_0 = [\beta_c, \beta_w, \beta_x] = [-0.2, -0.3, 0.5]$ and $\boldsymbol{\gamma}_0 = [\gamma_c, \gamma_x] = [0.1, 0.2]$. We generate n_{pool} observations and randomly split these n_{pool} observations into D equally-sized data sets, in which n_{pool} is selected at 500, 1000, 2000, and 5000, and D varies from 1 to 5. Note that $D = 1$ implies that we can simply apply the conventional MLE, IPW-MLE, and AIPW estimators without pooling $\boldsymbol{\beta}$ and τ_{ate}. The results for $D = 1$ serve as the benchmark to compare the results with other D. When D varies from 2 to 5, we apply our estimation and federated methods from Section

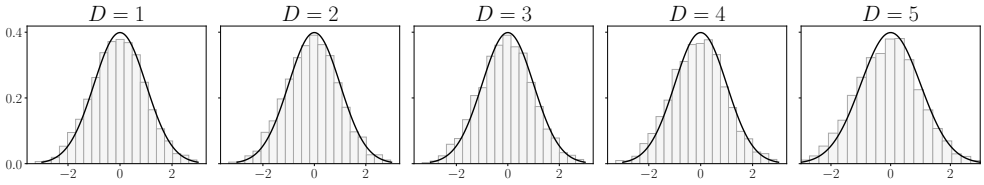
Figure 14: Histograms of Standardized MLE, IPW-MLE, and AIPW



(a) Federated MLE



(b) Federated IPW-MLE



(c) Federated AIPW Estimators

These figures show the histograms of estimated MLE, IPW-MLE, and AIPW estimators normalized by their estimated standard deviations, where $n_{\text{pool}} = 500$. D is selected from 1 to 5, where $D = 1$ implies that we estimate population parameters from fully pooled data and do not need to pool estimators. The normal density function is superimposed on the histograms. The results are based on 2,000 simulation replications.

3 to obtain the federated MLE, federated IPW-MLE, and federated AIPW estimators for $\boldsymbol{\beta}$ and τ_{ate} and their federated variances. We calculate the standardized federated MLE estimator using $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0)$ based on Theorem 1. Similarly, we calculate the standardized federated IPW-MLE and federated AIPW based on Theorems 2 and 3.

Figure 14 shows the histograms of standardized federated MLE and federated IPW-MLE for the treatment coefficient β_w, as well as federated AIPW for τ_{ate}, for various D with $n_{\text{pool}} = 500$ based on 2,000 replications of the above procedure. The histograms match the standard normal density function very well. Additionally, Table 13 reports the mean and standard error of the standardized federated MLE, federated IPW-MLE, and federated AIPW estimators for other n_{pool}. Figure 14 and Table 13 show that federated estimators across data sets are very close to those estimated from the combined, individual-level data. Moreover, they support the validity of our asymptotic results in finite samples even when n_{pool} is as low as 500. A sample size of a few hundred observations for good finite sample properties can be satisfied in many empirical medical applications, such as our medical claims data in Section 5.

Table 13: Simulations: Standardized Federated Maximum Likelihood Estimators

D	1		2		3		4		5	
n	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
500	-0.060	1.005	-0.049	1.000	-0.038	0.995	-0.027	0.987	-0.014	0.984
1000	-0.011	0.997	-0.004	0.994	0.004	0.991	0.010	0.988	0.018	0.986
2000	-0.035	0.999	-0.029	0.998	-0.025	0.997	-0.020	0.995	-0.013	0.993
5000	-0.015	1.019	-0.012	1.019	-0.008	1.018	-0.005	1.017	-0.002	1.017

(a) Federated MLE

D	1		2		3		4		5	
n	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
500	-0.058	1.004	-0.047	1.000	-0.036	0.994	-0.025	0.986	-0.012	0.983
1000	-0.012	0.996	-0.005	0.994	0.005	0.990	0.011	0.989	0.017	0.983
2000	-0.035	1.000	-0.030	0.999	-0.024	0.997	-0.019	0.998	-0.013	0.995
5000	-0.014	1.020	-0.011	1.019	-0.008	1.018	-0.005	1.018	-0.001	1.018

(b) Federated IPW-MLE

D	1		2		3		4		5	
n	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
500	-0.053	1.009	-0.060	1.014	-0.061	1.025	-0.071	1.036	-0.083	1.044
1000	-0.004	0.999	-0.008	1.003	-0.013	1.007	-0.015	1.014	-0.019	1.011
2000	-0.025	1.000	-0.029	1.002	-0.030	1.001	-0.034	1.007	-0.038	1.008
5000	-0.002	1.020	-0.005	1.020	-0.007	1.022	-0.009	1.022	-0.009	1.023

(c) Federated AIPW

This table reports the mean and standard error of the standardized federated MLE and federated IPW-MLE for the treatment coefficient β_w, as well as the standardized federated AIPW for ATE τ_{ate} across 2,000 simulation replications. n_{pool} is selected at 500, 1000, 2000, and 5000. D is selected from 1 to 5, where $D = 1$ implies that we estimate population parameters from the pooled data and do not need to pool estimators. The results for the federated estimators ($D = 2, 3, 4, 5$) are very close to those estimated from the pooled data ($D = 1$), implying the validity of our federated procedures for MLE, IPW-MLE, and AIPW. Moreover, the mean is close to 0, and the standard error is close to 1, verifying that our federated estimators have good finite sample properties.

Appendix E Proofs

Let $\dot{\boldsymbol{\ell}}_n(\boldsymbol{\beta}) \coloneqq \frac{\partial \boldsymbol{\ell}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ and $\ddot{\boldsymbol{\ell}}_n(\boldsymbol{\beta}) \coloneqq \frac{\partial^2 \boldsymbol{\ell}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}$ be the gradient and Hessian of the likelihood function. Moreover, let $\boldsymbol{\ell}_{n_k}^{(k)}(\boldsymbol{\beta})$, $\dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta})$, $\ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)}$ be the likelihood function, gradient, Hessian, and estimator on data set k.

E.1 Misspecified Maximum Likelihood Estimator

If the outcome model in the maximum likelihood estimator is misspecified, under suitable regularity conditions, the maximum likelihood estimator is still consistent and asymptotic normal (White, 1982), i.e.,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{mle}} - \boldsymbol{\beta}^*) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{A}_{\boldsymbol{\beta}^*}^{-1} \mathbf{B}_{\boldsymbol{\beta}^*} \mathbf{A}_{\boldsymbol{\beta}^*}^{-1}\big), \tag{26}$$

where $\boldsymbol{\beta}^*$ minimizes the Kullback-Leibler Information Criterion,

$$\int \log\Big(\frac{g(y|\mathbf{x}, w)}{f(y|\mathbf{x}, w, \boldsymbol{\beta})}\Big) dG(\mathbf{x}, w, y).$$

$G(\mathbf{x}, w, y)$ is the cumulative density function of $(\mathbf{x}, w, y)$. $g(y|\mathbf{x}, w)$ is the population density function of y given $(\mathbf{x}, w)$. $\mathbf{A}_{\boldsymbol{\beta}^*}$ and $\mathbf{B}_{\boldsymbol{\beta}^*}$ are $\mathbf{A}_{\boldsymbol{\beta}}$ and $\mathbf{B}_{\boldsymbol{\beta}}$ evaluated at $\boldsymbol{\beta}^*$ for the definitions of $\mathbf{A}_{\boldsymbol{\beta}}$ and $\mathbf{B}_{\boldsymbol{\beta}}$ provided in Table 1.

E.2 Proof of Proposition 1

Proof of Proposition 1. We adjust the covariates corresponding to $\boldsymbol{\beta}_{\mathcal{S}^\complement}^{(k)}$ by data set. For example, in generalized linear models, we can partition the treatment and covariates into two groups, $\tilde{\mathbf{X}}_i = (W_i, \mathbf{X}_i) = (\tilde{\mathbf{X}}_{\mathcal{S}}, \tilde{\mathbf{X}}_{\mathcal{S}^\complement})$, and include the interaction terms between $\tilde{\mathbf{X}}_{\mathcal{S}^\complement}$ and Z_k in the pooled outcome model, where Z_k is an binary variable indicating whether an observation is in data set k.

If Y_i follows a GLM, it means the conditional distribution of Y_i on $\mathbf{X}_i$ and W_i is in the exponential family and the log-likelihood function can be simplified to

$$\boldsymbol{\ell}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Y_i \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{\phi} + c(Y_i, \phi), \tag{27}$$

for a dispersion parameter ϕ, a natural parameter $\boldsymbol{\theta}$, and functions $b(\boldsymbol{\theta})$, and $c(Y, \phi)$.[24] Additionally, with link function g, we have $\mathbb{E}[Y_i] = \mu_i = b'(\boldsymbol{\theta}_i)$, $\tilde{\mathbf{X}}_i^\top \boldsymbol{\beta} = g(\mu_i)$ and $\tilde{\mathbf{X}}_i = (\mathbf{X}_i, W_i)$. Let $h(\tilde{\mathbf{X}}_i^\top \boldsymbol{\beta}) \coloneqq$

[24] By slight abuse of notation, $\boldsymbol{\ell}_n(\boldsymbol{\beta})$ is the shorthand for $\boldsymbol{\ell}_n(\boldsymbol{\beta}; \phi)$, and likewise for $\dot{\boldsymbol{\ell}}_n(\boldsymbol{\beta})$, $\ddot{\boldsymbol{\ell}}_n(\boldsymbol{\beta})$, and $\mathcal{I}(\boldsymbol{\beta})$ are similar.

$\boldsymbol{\theta}_i = (b')^{-1} \circ g^{-1}(\tilde{\mathbf{X}}_i^\top \boldsymbol{\beta})$. Therefore, we have $\dot{\boldsymbol{\ell}}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{Y_i - \mu_i}{\phi} h'(\tilde{\mathbf{X}}_i^\top \boldsymbol{\beta}) \tilde{\mathbf{X}}_i$ and

$$\mathbb{E}[\ddot{\boldsymbol{\ell}}_n(\boldsymbol{\beta})] = -\frac{1}{\phi} \sum_{i=1}^n b''(\boldsymbol{\theta}_i)[h'(\tilde{\mathbf{X}}_i^\top \boldsymbol{\beta})]^2 \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^\top = -\sum_{i=1}^n \underbrace{\frac{h'(\tilde{\mathbf{X}}_i^\top \boldsymbol{\beta})}{g'(\mu_i)\phi}}_{\xi_i} \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^\top = -\tilde{\mathbf{X}}^\top \Xi \tilde{\mathbf{X}},$$

where $\Xi = \text{diag}(\xi_1, \cdots, \xi_n)$. We have $\mathcal{I}(\boldsymbol{\beta}) = \tilde{\mathbf{X}}^\top \Xi \tilde{\mathbf{X}}$ and $\text{Var}(\hat{\boldsymbol{\beta}}) = (\tilde{\mathbf{X}}^\top \Xi \tilde{\mathbf{X}})^{-1}$.

Now we consider two data sets with parameters $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$.

Suppose $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$ but we use a richer model for the pooled data that adjusts covariates by data sets, $(\tilde{\mathbf{X}}_{i,\mathcal{S}}, \tilde{\mathbf{X}}_{i,\mathcal{S}^\complement} \cdot Z_1, \tilde{\mathbf{X}}_{i,\mathcal{S}^\complement} \cdot Z_2)$, with coefficients $(\boldsymbol{\beta}_\mathcal{S}, \boldsymbol{\beta}_{\mathcal{S}^\complement}^{(1)}, \boldsymbol{\beta}_{\mathcal{S}^\complement}^{(2)})$, where Z_1 and Z_2 are binary variables indicating whether an observation is in data sets 1 and 2, respectively. We show that using this richer model gives us a less efficient estimate of $\boldsymbol{\beta}_\mathcal{S}$, where the estimator is denoted as $\boldsymbol{\beta}_\mathcal{S}^{\text{sep}}$. The corresponding estimate of $\boldsymbol{\beta}$ from the simple model is denoted as $\boldsymbol{\beta}_\mathcal{S}^{\text{joint}}$.

Next we show

$$\text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{sep}}) \succcurlyeq \text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{joint}}).$$

Let $\tilde{\mathbf{X}}_\mathcal{S}^{(j)} \in \mathbb{R}^{n_j \times |\mathcal{S}|}$ and $\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(j)} \in \mathbb{R}^{n_j \times (d - |\mathcal{S}|)}$ be the covariate matrices of shared parameters and dataset-specific parameters on data set j. With algebra, we can show that

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_w^{\text{sep}})^{-1} &= (\tilde{\mathbf{X}}_\mathcal{S}^{(1)})^\top (\Xi^{(1)})^{-1} \tilde{\mathbf{X}}_\mathcal{S}^{(1)} - \big((\tilde{\mathbf{X}}_\mathcal{S}^{(1)})^\top (\Xi^{(1)})^{-1} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)}\big) \cdot \big((\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)})^\top (\Xi^{(1)})^{-1} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)}\big)^{-1} \cdot \big((\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)})^\top (\Xi^{(1)})^{-1} \tilde{\mathbf{X}}_\mathcal{S}^{(1)}\big) \\
&\quad + (\tilde{\mathbf{X}}_\mathcal{S}^{(2)})^\top (\Xi^{(2)})^{-1} \tilde{\mathbf{X}}_\mathcal{S}^{(2)} - \big((\tilde{\mathbf{X}}_\mathcal{S}^{(2)})^\top (\Xi^{(2)})^{-1} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)}\big) \cdot \big((\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)})^\top (\Xi^{(2)})^{-1} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)}\big)^{-1} \cdot \big((\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)})^\top (\Xi^{(2)})^{-1} \tilde{\mathbf{X}}_\mathcal{S}^{(2)}\big) \\
\text{Var}(\hat{\boldsymbol{\beta}}_w^{\text{joint}})^{-1} &= \big((\tilde{\mathbf{X}}_\mathcal{S}^{(1)})^\top (\Xi^{(1)})^{-1} \tilde{\mathbf{X}}_\mathcal{S}^{(1)} + (\tilde{\mathbf{X}}_\mathcal{S}^{(2)})^\top (\Xi^{(2)})^{-1} \tilde{\mathbf{X}}_\mathcal{S}^{(2)}\big) - \big((\tilde{\mathbf{X}}_\mathcal{S}^{(1)})^\top (\Xi^{(1)})^{-1} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)} + (\tilde{\mathbf{X}}_\mathcal{S}^{(2)})^\top (\Xi^{(2)})^{-1} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)}\big) \\
&\quad \cdot \big((\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)})^\top (\Xi^{(1)})^{-1} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)} + (\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)})^\top (\Xi^{(2)})^{-1} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)}\big)^{-1} \cdot \big((\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)})^\top (\Xi^{(1)})^{-1} \tilde{\mathbf{X}}_\mathcal{S}^{(1)} + (\tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)})^\top (\Xi^{(2)})^{-1} \tilde{\mathbf{X}}_\mathcal{S}^{(2)}\big).
\end{aligned}$$

In order to show $\text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{sep}}) \succcurlyeq \text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{joint}})$, it is equivalent to show $\text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{sep}})^{-1} \preccurlyeq \text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{joint}})^{-1}$ and therefore equivalent to show for any vector $v \in R^{|\mathcal{S}|}$, $v^\top \text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{sep}})^{-1} v \leq v^\top \text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{joint}})^{-1} v$. Let $a_1 = (\Xi^{(1)})^{-1/2} \tilde{\mathbf{X}}_\mathcal{S}^{(1)} \cdot v$, $a_2 = (\Xi^{(2)})^{-1/2} \tilde{\mathbf{X}}_\mathcal{S}^{(2)} \cdot v$, $\mathbf{M}_1 = (\Xi^{(1)})^{-1/2} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(1)}$ and $\mathbf{M}_2 = (\Xi^{(2)})^{-1/2} \tilde{\mathbf{X}}_{\mathcal{S}^\complement}^{(2)}$. With algebra, we have

$$\begin{aligned}
&v^\top \text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{sep}})^{-1} v \leq v^\top \text{Var}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{\text{joint}})^{-1} v \qquad (28) \\
\Leftrightarrow &(a_1)^\top a_1 - (a_1)^\top \mathbf{M}_1 \big((\mathbf{M}_1)^\top \mathbf{M}_1\big)^{-1} (\mathbf{M}_1)^\top a_1 + (a_2)^\top a_2 - (a_2)^\top \mathbf{M}_2 \big((\mathbf{M}_2)^\top \mathbf{M}_2\big)^{-1} (\mathbf{M}_2)^\top a_2 \\
&\leq (a_1)^\top a_1 + (a_2)^\top a_2 - \big((a_1)^\top \mathbf{M}_1 + (a_2)^\top \mathbf{M}_2\big) \big((\mathbf{M}_1)^\top \mathbf{M}_1 + (\mathbf{M}_2)^\top \mathbf{M}_2\big)^{-1} \big((\mathbf{M}_1)^\top a_1 + (\mathbf{M}_2)^\top a_2\big).
\end{aligned}$$

Consider the SVD of $\mathbf{M}_1 = \mathbf{U}_1 \mathbf{D}_1 \mathbf{V}_1^\top \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{M}_2 = \mathbf{U}_2 \mathbf{D}_2 \mathbf{V}_2^\top \in \mathbb{R}^{n_2 \times p}$, where $\mathbf{V}_1^{-1} = \mathbf{V}_1^\top$ and $\mathbf{V}_2^{-1} = \mathbf{V}_2^\top$ following $p \ll n_1$ and $p \ll n_2$. We can simplify the inequality (28) to

$$\begin{aligned}
&a_1^\top \mathbf{U}_1 \mathbf{U}_1^\top a_1 + a_2^\top \mathbf{U}_2 \mathbf{U}_2^\top a_2 \\
\geq &\big(a_1^\top \mathbf{U}_1 \mathbf{D}_1 \mathbf{V}_1^\top + a_2^\top \mathbf{U}_2 \mathbf{D}_2 \mathbf{V}_2^\top\big) \big(\mathbf{V}_1 \mathbf{D}_1^2 \mathbf{V}_1^\top + \mathbf{V}_2 \mathbf{D}_2^2 \mathbf{V}_2^\top\big)^{-1} \big(\mathbf{V}_1 \mathbf{D}_1 \mathbf{U}_1^\top a_1 + \mathbf{V}_2 \mathbf{D}_2 \mathbf{U}_2^\top a_2\big).
\end{aligned}$$

Let $\mathbf{\Omega} = \mathbf{D}_1^{-1}\mathbf{V}_1^{-1}\mathbf{V}_2\mathbf{D}_2$. We can write the terms in the above in equality as functions of $\mathbf{\Omega}$:

$$\begin{aligned}
\mathbf{D}_1\mathbf{V}_1^\top\big(\mathbf{V}_1\mathbf{D}_1^2\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2^\top\big)^{-1}\mathbf{V}_1\mathbf{D}_1 =& \big(\mathbf{I} + \mathbf{D}_1^{-1}\mathbf{V}_1^{-1}\mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2^\top(\mathbf{V}_1^\top)^{-1}\mathbf{D}_1^{-1}\big)^{-1} = \big(\mathbf{I} + \mathbf{\Omega}\mathbf{\Omega}^\top\big)^{-1} \\
\mathbf{D}_2\mathbf{V}_2^\top\big(\mathbf{V}_1\mathbf{D}_1^2\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2^\top\big)^{-1}\mathbf{V}_2\mathbf{D}_2 =& \big(\mathbf{I} + \mathbf{D}_2^{-1}\mathbf{V}_2^{-1}\mathbf{V}_1\mathbf{D}_1^2\mathbf{V}_1^\top(\mathbf{V}_2^\top)^{-1}\mathbf{D}_2^{-1}\big)^{-1} = \big(\mathbf{I} + \mathbf{\Omega}^{-1}(\mathbf{\Omega}^{-1})^\top\big)^{-1} \\
\mathbf{D}_1\mathbf{V}_1^\top\big(\mathbf{V}_1\mathbf{D}_1^2\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{D}_2^2\mathbf{V}_2^\top\big)^{-1}\mathbf{V}_2\mathbf{D}_2 =& (\mathbf{\Omega}^{-1} + \mathbf{\Omega}^\top)^{-1}.
\end{aligned}$$

Let $\tilde{a}_1 = \mathbf{U}_1^\top a_1$ and $\tilde{a}_2 = \mathbf{U}_2^\top a_2$,. We can further simplify Inequality (28) to

$$\tilde{a}_1^\top\tilde{a}_1 + \tilde{a}_2^\top\tilde{a}_2 \geq \tilde{a}_1^\top\big(\mathbf{I} + \mathbf{\Omega}\mathbf{\Omega}^\top\big)^{-1}\tilde{a}_1 + \tilde{a}_2^\top\big(\mathbf{I} + \mathbf{\Omega}^{-1}(\mathbf{\Omega}^{-1})^\top\big)^{-1}\tilde{a}_2 + 2a_1(\mathbf{\Omega}^{-1} + \mathbf{\Omega}^\top)^{-1}a_2.$$

Consider the SVD of $\mathbf{\Omega} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. We can further simplify Inequality (28) to

$$\tilde{a}_1^\top\mathbf{U}\mathbf{D}^2\big(\mathbf{I} + \mathbf{D}^2\big)^{-1}\mathbf{U}^\top\tilde{a}_1 + \tilde{a}_2^\top\mathbf{V}\mathbf{D}^{-2}\big(I + \mathbf{D}^{-2}\big)^{-1}\mathbf{V}^{-1}\tilde{a}_2 \geq 2\tilde{a}_1^\top\mathbf{U}\big(\mathbf{D} + \mathbf{D}^{-1}\big)^{-1}\mathbf{V}^\top\tilde{a}_2$$

Denote each element in $\mathbf{U}^\top\tilde{a}_1$ as $\bar{a}_{1,i}$ and each element in $\mathbf{V}^{-1}\tilde{a}_2$ as $\bar{a}_{2,i}$. We can further simplify Inequality (28) to

$$\sum_i \frac{\bar{a}_{1,i}^2 d_i^2}{1 + d_i^2} + \sum_i \frac{\bar{a}_{2,i}^2}{1 + d_i^2} \geq 2\sum_i \frac{\bar{a}_{1,i}\bar{a}_{2,i}d_i}{1 + d_i^2}$$

We can see that this inequality holds from the Cauchy-Schwarz inequality, and therefore Inequality (28) holds. If there are more data sets, $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{sep}}) \succcurlyeq \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{joint}})$ still holds by induction.

□

E.3 Proof of Results for Federated MLE in Section 4.1

E.3.1 Proof of Theorem 1

Proof of Theorem 1. Our proof of Theorem 1 consists of showing the following four equations:

1. $n_{\mathrm{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\mathrm{pool}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{mle}}^{\mathrm{pool}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$
2. $n_{\mathrm{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\mathrm{fed}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{mle}}^{\mathrm{fed}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$
3. $n_{\mathrm{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\mathrm{pool}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{mle}}^{\mathrm{fed}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$
4. $n_{\mathrm{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\mathrm{fed}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{mle}}^{\mathrm{pool}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$

We need to consider two cases. The first case is the information matrix $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ being the same for all data sets. The second case is the information matrix varying across data sets. The first case consists of two sub-cases depending on whether Assumption 2 holds or not.

The first step is to show $n_{\mathrm{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\mathrm{pool}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{mle}}^{\mathrm{pool}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$. MLE is consistent and

asymptotic normal (see Chapter 4.2.3 in Amemiya (1985)):

$$\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} \xrightarrow{p} \boldsymbol{\beta}_0$$
$$\sqrt{n_k}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0\big) \xrightarrow{d} \mathcal{N}\big(0, \mathcal{I}^{(k)}(\boldsymbol{\beta}_0)^{-1}\big),$$

where $\mathcal{I}^{(k)}(\boldsymbol{\beta}) = -\mathbb{E}_{(\mathbf{x},w,y)\sim\mathbb{P}^{(k)}}\Big[\frac{\partial^2 \log f(y_i|\mathbf{x}_i,w_i,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\Big]$. From the law of large numbers and the consistency of $\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)}$, we have $-\frac{1}{n_k}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} \xrightarrow{p} \mathcal{I}^{(k)}(\boldsymbol{\beta}_0)$. Hence, from Slutsky's theorem, we have for each individual data set k,

$$\sqrt{n_k}\big(-\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}/n_k\big)^{-1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0\big) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{I}_d\big),$$

where $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} = \ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)})$. Similarly for the combined, individual-level data, we have $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}} \xrightarrow{p} \boldsymbol{\beta}_0$ and $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{pool}} = \big(-\frac{1}{n_{\text{pool}}}\sum_{k=1}^{D}\ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}})\big)^{-1} \xrightarrow{p} \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1}$. Then, we have

$$n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{pool}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d),$$

which is our first equation.

The second step is to show $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$.

Let us first consider the case where the information matrix $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ is the same for all data sets (and then follow with the case where $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ differs across data sets). Let $\mathcal{I}(\boldsymbol{\beta}) = \mathcal{I}^{(k)}(\boldsymbol{\beta})$ for all k. In this case, $\mathcal{I}(\boldsymbol{\beta}) = \mathcal{I}^{\text{pool}}(\boldsymbol{\beta})$. Using the property that for all k, $-\frac{1}{n_k}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} \xrightarrow{p} \mathcal{I}(\boldsymbol{\beta}_0)$, we have

$$\Big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\Big)^{-1} \cdot \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(j)} \cdot \frac{\sum_{k=1}^{D} n_k}{n_j} \xrightarrow{p} \mathbf{I}_d, \tag{29}$$

and we can use this property to show the consistency of $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}}$. Let $\hat{p}_{n,j} = \frac{n_j}{\sum_{k=1}^{D} n_k}$. We have

$$\begin{aligned}
\Big\|\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0\Big\|_2 &= \Bigg\|\Big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\Big)^{-1}\Big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0\big)\Big)\Bigg\|_2 \\
&= \Bigg\|\sum_{j=1}^{D}\hat{p}_{n,j}\cdot\Big[\Big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\Big)^{-1}\cdot\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(j)}\cdot\frac{1}{\hat{p}_{n,j}}\cdot\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(j)} - \boldsymbol{\beta}_0\big)\Big]\Bigg\|_2 \\
&\le \sum_{j=1}^{D}\hat{p}_{n,j}\cdot\underbrace{\Bigg\|\Big[\Big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\Big)^{-1}\cdot\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(j)}\cdot\frac{1}{\hat{p}_{n,j}}\cdot\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(j)} - \boldsymbol{\beta}_0\big)\Big]\Bigg\|_2}_{o_p(1)} = o_p(1),
\end{aligned} \tag{30}$$

where we use the properties that $\hat{\boldsymbol{\beta}}_{\text{mle}}^{(j)} \xrightarrow{p} \boldsymbol{\beta}_0$, $0 < \hat{p}_{n,j} < 1$ and D is finite.

Since observations between data sets are asymptotically independent, we have $\Big(n_1^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(1)} - \boldsymbol{\beta}_0\big), n_2^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(2)} - \boldsymbol{\beta}_0\big), \cdots, n_D^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(D)} - \boldsymbol{\beta}_0\big)\Big)$ jointly converge to a normal distribution, and for any

$j \neq k$, $n_j^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(j)} - \boldsymbol{\beta}_0\big)$ and $n_k^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0\big)$ are independent. Using $n_{\text{pool}} = \sum_{k=1}^{D} n_k$, we can decompose $n_{\text{pool}}^{1/2}\Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0\Big)$ as

$$n_{\text{pool}}^{1/2}\Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0\Big) = \sum_{j=1}^{D} \hat{p}_{n,j}^{1/2} \underbrace{\Big[\Big(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\Big)^{-1} \cdot \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(j)} \cdot \frac{1}{\hat{p}_{n,j}} \cdot n_j^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(j)} - \boldsymbol{\beta}_0\big)\Big]}_{:=\boldsymbol{\xi}_{n_j}^{(j)}}.$$

For the term $\boldsymbol{\xi}_{n_j}^{(j)}$ in the bracket, from Eq. (29) and Slutsky's theorem, we have

$$\boldsymbol{\xi}_{n_j}^{(j)} \xrightarrow{d} \boldsymbol{\xi}^{j} \overset{d}{=} \mathcal{N}(0, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}).$$

Let us consider the multiplier $\hat{p}_{n,j}^{1/2}$. If Assumption 2 holds, then this multiplier converges, i.e., $\hat{p}_{n,j}^{1/2} \to p_j^{1/2}$, where p_j is defined in Assumption 2. Therefore, from Slutsky's theorem and the delta method, we have

$$n_{\text{pool}}^{1/2}\Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0\Big) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}).$$

A more challenging case is when Assumption 2 is violated. As a result, $\hat{p}_{n,j}$ may not be a convergent sequence. We need to use the almost sure representation theory (Pollard, 2012). From the joint convergence of $\boldsymbol{\xi}_{n_j}^{(j)}$, i.e., $(\boldsymbol{\xi}_{n_1}^{(1)}, \cdots, \boldsymbol{\xi}_{n_D}^{(D)}) \xrightarrow{d} (\boldsymbol{\xi}^{(1)}, \cdots, \boldsymbol{\xi}^{(D)})$, the almost sure representation theory asserts that there exist random vectors $(\boldsymbol{\xi}_{n_1}^{(1)*}, \cdots, \boldsymbol{\xi}_{n_D}^{(D)*})$ and $(\boldsymbol{\xi}^{(1)*}, \cdots, \boldsymbol{\xi}^{(D)*})$ with the same distributions as $(\boldsymbol{\xi}_{n_1}^{(1)}, \cdots, \boldsymbol{\xi}_{n_D}^{(D)})$ and $(\boldsymbol{\xi}^{(1)}, \cdots, \boldsymbol{\xi}^{(D)})$, respectively, such that $(\boldsymbol{\xi}_{n_1}^{(1)*}, \cdots, \boldsymbol{\xi}_{n_D}^{(D)*}) \xrightarrow{a.s.} (\boldsymbol{\xi}^{(1)*}, \cdots, \boldsymbol{\xi}^{(D)*})$. Then we have the following decomposition:

$$\sum_{j=1}^{D} \hat{p}_{n,j}^{1/2} \cdot \boldsymbol{\xi}_{n_j}^{(j)*} = \sum_{j=1}^{D} \hat{p}_{n,j}^{1/2} \cdot \boldsymbol{\xi}^{(j)*} + \sum_{j=1}^{D} \hat{p}_{n,j}^{1/2} \cdot \underbrace{\big(\boldsymbol{\xi}_{n_j}^{(j)*} - \boldsymbol{\xi}^{(j)*}\big)}_{\xrightarrow{a.s.} 0}.$$

For the first term in this decomposition, we have

$$\sum_{j=1}^{D} \hat{p}_{n,j}^{1/2} \cdot \boldsymbol{\xi}^{(j)*} \overset{d}{=} \mathcal{N}\Big(0, \sum_{j=1}^{D} \hat{p}_{n,j} \cdot \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\Big) \overset{d}{=} \mathcal{N}\big(0, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big),$$

using the properties that $\boldsymbol{\xi}^{(j)*}$ is independent of $\boldsymbol{\xi}^{(k)*}$, $\boldsymbol{\xi}^{(j)*}$ follows a normal distribution with variance $\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$, and $\sum_{j=1}^{D} \hat{p}_{n,j} = 1$. Left multiplying $\mathcal{I}(\boldsymbol{\beta}_0)^{1/2}$ in the above equation, we have

$$\mathcal{I}(\boldsymbol{\beta}_0)^{1/2} \sum_{j=1}^{D} \hat{p}_{n,j}^{1/2} \cdot \boldsymbol{\xi}_{n_j}^{(j)*} \xrightarrow{d} \mathcal{N}\big(0, \mathbf{I}_d\big).$$

Using the property that $(\boldsymbol{\xi}_{n_1}^{(1)*}, \cdots, \boldsymbol{\xi}_{n_D}^{(D)*})$ and $(\boldsymbol{\xi}_{n_1}^{(1)}, \cdots, \boldsymbol{\xi}_{n_D}^{(D)})$ have the same distribution, we have

$$n_{\text{pool}}^{1/2} \cdot \mathcal{I}(\boldsymbol{\beta}_0)^{1/2} \cdot \Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0\Big) \overset{d}{=} \mathcal{I}(\boldsymbol{\beta}_0)^{1/2} \sum_{j=1}^{D} \hat{p}_{n,j}^{1/2} \cdot \boldsymbol{\xi}_{n_j}^{(j)} \overset{d}{\to} \mathcal{N}\big(0, \mathbf{I}_d\big).$$

Using the property that $\big(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}}\big)^{-1} = \sum_{j=1}^{D} \hat{p}_{n,j} \cdot \big(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{(j)}\big)^{-1} \overset{p}{\to} \mathcal{I}(\boldsymbol{\beta}_0)$ (where we obtain $\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}}$ by sample size weighting), from Slutsky's theorem we have

$$n_{\text{pool}}^{1/2} (\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}})^{-1/2} (\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0) \overset{d}{\to} \mathcal{N}\big(0, \mathbf{I}_d\big). \tag{31}$$

This recovers our second equation in the case where $\mathcal{I}^{(k)}(\boldsymbol{\beta}_0)$ is the same for all k.

Next, let us consider the case, where $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ varies with the data set. In this case, we assume Assumption 2 holds. Using the property that $\frac{1}{n_k} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} \overset{p}{\to} -\mathcal{I}^{(k)}(\boldsymbol{\beta}_0)$ and the definition $\mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0) = \sum_{k=1}^{D} p_k \mathcal{I}^{(k)}(\boldsymbol{\beta}_0)$, we have

$$\frac{1}{\sum_{k=1}^{D} n_k} \sum_{j=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(j)} = -\sum_{j=1}^{D} \underbrace{\frac{n_j}{\sum_{k=1}^{D} n_k}}_{\hat{p}_{n,j}} \mathcal{I}^{(j)}(\boldsymbol{\beta}_0) + o_p(1) = -\mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0) + o_p(1).$$

Since $\big\|\mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1} \cdot \mathcal{I}^{(j)}(\boldsymbol{\beta}_0)\big\|_2 \leq M$, we have

$$\Big(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\Big)^{-1} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(j)} \cdot \frac{1}{\hat{p}_{n,j}} \overset{p}{\to} \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1} \cdot \mathcal{I}^{(j)}(\boldsymbol{\beta}_0).$$

and we can show the consistency of $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}}$ following the same procedures as Inequality (30) using the property that $\big\|\mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1} \cdot \mathcal{I}^{(j)}(\boldsymbol{\beta}_0)\big\|_2 \leq M$. For the asymptotic normality of $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}}$, since $\frac{n_j}{\sum_{k=1}^{D} n_k}$ converges to some constant for all j, using Slutsky's Theorem and the delta method, we have

$$\begin{aligned}
&n_{\text{pool}}^{1/2}\Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0\Big) = \sum_{j=1}^{D} \hat{p}_{n,j}^{1/2} \Big[\Big(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\Big)^{-1} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(j)} \cdot \frac{1}{\hat{p}_{n,j}} \cdot n_j^{1/2} \big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(j)} - \boldsymbol{\beta}_0\big)\Big] \\
\overset{d}{\to} &\mathcal{N}\Big(0, \sum_{j=1}^{D} p_j \cdot \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1} \cdot \mathcal{I}^{(j)}(\boldsymbol{\beta}_0) \cdot \mathcal{I}^{(j)}(\boldsymbol{\beta}_0)^{-1} \cdot \mathcal{I}^{(j)}(\boldsymbol{\beta}_0) \cdot \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1}\Big) \overset{d}{=} \mathcal{N}\big(0, \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1}\big).
\end{aligned}$$

Using the property $\big(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}}\big)^{-1} = \sum_{j=1}^{D} \hat{p}_{n,j} \cdot \big(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{(j)}\big)^{-1} = \sum_{j=1}^{D} p_j \mathcal{I}^{(j)}(\boldsymbol{\beta}_0) + o_p(1) = \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0) + o_p(1)$, (31) continues to hold, and we finish showing the second step for the case where $\mathcal{I}^{(k)}(\boldsymbol{\beta}_0)$ varies with k.

The third step is to show $n_{\text{pool}}^{1/2} (\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{pool}})^{-1/2} (\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0) \overset{d}{\to} \mathcal{N}(0, \mathbf{I}_d)$. From the second step, we have shown that $n_{\text{pool}}^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0\big) \overset{d}{\to} \mathcal{N}\big(0, \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1}\big)$ holds regardless of whether $\mathcal{I}^{(k)}(\boldsymbol{\beta}_0)$ varies

with k. From the first step, we have $\frac{1}{n_{\text{pool}}}\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{pool}} \xrightarrow{p} \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0)^{-1}$. By Slutsky's theorem,

$$n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{pool}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$$

which completes the proof of the third step.

The last step is to show $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$. We have shown that $\big(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}}\big)^{-1} = \mathcal{I}^{\text{pool}}(\boldsymbol{\beta}_0) + o_p(1)$ in the second step. Using this property, together with the first step, we have

$$n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_{\boldsymbol{\beta}}^{\text{fed}})^{-1/2}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d).$$

This recovers all four steps and therefore concludes the proof of Theorem 1.

□

E.3.2 Proof of Proposition 2

Proof of Proposition 2. We can show Proposition 2 following exactly the same steps as Theorem 1, but with $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ replaced by $(\mathbf{A}_{\boldsymbol{\beta}}^{(k)})^{-1}\mathbf{B}_{\boldsymbol{\beta}}^{(k)}(\mathbf{A}_{\boldsymbol{\beta}}^{(k)})^{-1}$, where the definitions of $\mathbf{A}_{\boldsymbol{\beta}}^{(k)}$ and $\mathbf{B}_{\boldsymbol{\beta}}^{(k)}$ can be found in Table 1. □

E.3.3 Proof of Proposition 3

Proof of Proposition 3. Since the nonzero blocks $\mathbf{A}_{\boldsymbol{\beta}_{\mathcal{S}},\boldsymbol{\beta}_{\mathcal{S}}}^{(k)}$, $\mathbf{A}_{\boldsymbol{\beta}_{\mathcal{S}},\boldsymbol{\beta}_{\mathcal{S}^{\complement}}^{(k)}}^{(k)}$, and $\mathbf{A}_{\boldsymbol{\beta}_{\mathcal{S}^{\complement}}^{(k)},\boldsymbol{\beta}_{\mathcal{S}^{\complement}}^{(k)}}^{(k)}$ in $\mathbf{A}^{\text{pad},(k)}$ can be consistently estimated, $\mathbf{A}^{\text{pad},(k)}$ can be consistently estimated for all k. Under Assumption 2, p_k can be consistently estimated. Hence, our pooling procedure provides a consistent estimator for $\mathbf{A}^{\text{pool}}$ (and similarly for $\mathbf{B}^{\text{pool}}$), where $\mathbf{A}^{\text{pool}}$ is defined as $\mathbf{A}^{\text{pool}} = \sum_{k=1}^{D} p_k \mathbf{A}^{\text{pad},(k)}$ (and $\mathbf{B}^{\text{pool}}$ is defined similarly). In the case where the outcome model is correctly specified (Condition 8), $\mathbf{A}^{\text{pool}} = \mathbf{B}^{\text{pool}}$.

Let $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}}$ be the estimator that maximizes the likelihood function $\boldsymbol{\ell}_{n_{\text{pool}}}^{\text{pool}}(\boldsymbol{\beta}^{\text{pool}})$ for the combined, individual-level data, where the true parameter is $\boldsymbol{\beta}_0^{\text{pool}}$. We have

$$n_{\text{pool}}^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}} - \boldsymbol{\beta}_0^{\text{pool}}\big) \xrightarrow{d} \mathcal{N}\big(0, (\mathbf{A}^{\text{pool}})^{-1}\mathbf{B}^{\text{pool}}(\mathbf{A}^{\text{pool}})^{-1}\big).$$

Recall that $\sqrt{n_k}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0^{(k)}\big) \xrightarrow{d} \mathcal{N}\big(0, (\mathbf{A}^{(k)})^{-1}\mathbf{B}^{(k)}(\mathbf{A}^{(k)})^{-1}\big)$ and $-\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}/n_k \xrightarrow{p} \mathbf{A}^{(k)}$. From Slutsky's theorem, we have $n_k^{-1/2}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0^{(k)}\big) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{B}^{(k)}\big)$, and then we have

$$n_k^{-1/2}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pad},(k)} - \boldsymbol{\beta}_0^{\text{pad},(k)}\big) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{B}^{\text{pad},(k)}\big),$$

using the property that $-\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}/n_k \xrightarrow{p} \mathbf{A}^{\text{pad},(k)}$. Moreover, we have

$$n_{\text{pool}} \cdot n_k^{-1/2}\Big(\sum_{j=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(j)}\Big)^{-1}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pad},(k)} - \boldsymbol{\beta}_0^{\text{pad},(k)}\big) \xrightarrow{d} \mathcal{N}\big(0, (\mathbf{A}^{\text{pool}})^{-1}\mathbf{B}^{\text{pad},(k)}(\mathbf{A}^{\text{pool}})^{-1}\big),$$

which follows from $-\frac{1}{n_{\text{pool}}}\sum_{j=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(j)} = -\sum_{j=1}^{D}\frac{n_j}{n_{\text{pool}}}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(j)}/n_j \xrightarrow{p} \sum_{j=1}^{D} p_j \mathbf{A}^{\text{pad},(j)} = \mathbf{A}^{\text{pool}}$ from Assumption 2.

Note that we have the equality that $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}\boldsymbol{\beta}_0^{\text{pad},(k)} = \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}\boldsymbol{\beta}_0^{\text{pool}}$. This equality follows from the fact that for all the nozero entries in $\boldsymbol{\beta}_0^{\text{pool}} - \boldsymbol{\beta}_0^{\text{pad},(k)}$, the corresponding columns in $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}$ are 0. Then, we can decompose $\boldsymbol{\beta}_0^{\text{pool}}$ as

$$\boldsymbol{\beta}_0^{\text{pool}} = \sum_{k=1}^{D}\Big(\sum_{j=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(j)}\Big)^{-1}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}\boldsymbol{\beta}_0^{\text{pad},(k)}.$$

Now we are ready to show the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}}$:

$$n_{\text{pool}}^{1/2}\Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0^{\text{pool}}\Big) = \sum_{k=1}^{D}\frac{n_k^{1/2}}{n_{\text{pool}}^{1/2}}\frac{n_{\text{pool}}}{n_k^{1/2}}\Big(\sum_{j=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(j)}\Big)^{-1}\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{\text{pad},(k)}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pad},(k)} - \boldsymbol{\beta}_0^{\text{pad},(k)}\big)$$
$$\xrightarrow{d}\mathcal{N}\Big(0, (\mathbf{A}^{\text{pool}})^{-1}\Big(\sum_{k=1}^{D}p_k\mathbf{B}^{\text{pad},(k)}\Big)(\mathbf{A}^{\text{pool}})^{-1}\Big) \overset{d}{=} \mathcal{N}\Big(0, (\mathbf{A}^{\text{pool}})^{-1}\mathbf{B}^{\text{pool}}(\mathbf{A}^{\text{pool}})^{-1}\Big).$$

Hence, we have $n_{\text{pool}}^{1/2}\Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0^{\text{pool}}\Big) \overset{d}{=} n_{\text{pool}}^{1/2}\Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{pool}} - \boldsymbol{\beta}_0^{\text{pool}}\Big)$. Our pooling procedures provide consistent estimators for $\mathbf{A}^{\text{pool}}$ and $\mathbf{B}^{\text{pool}}$. Then, we follow the same procedures and can show that the four steps in the proof of Theorem 1 continue to hold (even with a misspecified outcome model). □

E.3.4 Proof of Proposition 5

Proof of Proposition 5. Since we have model shift, the MLE estimator satisfies

$$\sqrt{n_k}\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0^{(k)}\big) \xrightarrow{d}\mathcal{N}\Big(0, \mathcal{I}^{(k)}(\boldsymbol{\beta}_0)^{-1}\Big).$$

In this proof, let $\mathbf{H}^{(k)}(\boldsymbol{\beta}) = \sum_{i=1}^{n_k}\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\log f(Y_i^{(k)}|\mathbf{X}_i^{(k)}, W_i^{(k)}, \boldsymbol{\beta})$. From the mean value theorem, on each data set k, we have

$$\Big(\frac{1}{n_k}\ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)})\Big)\big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0^{(k)}\big) = -\frac{1}{n_k}\dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}_0^{(k)}) + o_p\Big(\frac{1}{\sqrt{n_k}}\Big),$$

and the above equation holds with $\ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)})$ replaced by $\ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}_0^{(k)})$. Since $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} = \ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)})$ for all k, we have

$$
\begin{aligned}
&\frac{1}{n_{\text{pool}}} \sum_{k=1}^{D} \Big(\ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)}) \big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}^* \big) \Big) \\
=& - \frac{1}{n_{\text{pool}}} \sum_{k=1}^{D} \Big(\dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}_0^{(k)}) - \ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}_0^{(k)}) \big(\boldsymbol{\beta}_0^{(k)} - \boldsymbol{\beta}^* \big) \Big) + o_p\big(n_{\text{pool}}^{-1/2} \big) \\
=& - \frac{1}{n_{\text{pool}}} \sum_{k=1}^{D} \big(\dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}_0^{(k)}) - \ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}_0^{(k)}) \cdot \boldsymbol{\beta}_0^{(k)} \big) - \Big(\frac{1}{n_{\text{pool}}} \sum_{j=1}^{D} \ddot{\boldsymbol{\ell}}_{n_j}^{(j)}(\boldsymbol{\beta}^*) \Big) \boldsymbol{\beta}^* + o_p\big(n_{\text{pool}}^{-1/2} \big) \\
=& - \frac{1}{n_{\text{pool}}} \sum_{k=1}^{D} \big(\dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}^*) - \ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}^*) \cdot \boldsymbol{\beta}^* \big) - \Big(\frac{1}{n_{\text{pool}}} \sum_{j=1}^{D} \ddot{\boldsymbol{\ell}}_{n_j}^{(j)}(\boldsymbol{\beta}^*) \Big) \boldsymbol{\beta}^* + o_p\big(n_{\text{pool}}^{-1/2} \big) \\
=& - \frac{1}{n_{\text{pool}}} \sum_{k=1}^{D} \dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}^*) + o_p\big(n_{\text{pool}}^{-1/2} \big),
\end{aligned}
$$

where the first equality follows from that $p_k = \lim n_k / n_{\text{pool}}$ is bounded away from 0 and 1, the second equality follows from the assumption that $\mathcal{I}^{(j)}(\boldsymbol{\beta})$ not depending on $\boldsymbol{\beta}$ (recall $\ddot{\boldsymbol{\ell}}_{n_j}^{(j)}(\boldsymbol{\beta})/n_j \xrightarrow{p} \mathcal{I}^{(j)}(\boldsymbol{\beta})$), and the third equality follows from the assumption that $\dot{\mathbf{d}}_y^{(j)}(\boldsymbol{\beta}) - \mathcal{I}^{(j)}(\boldsymbol{\beta}) \cdot \boldsymbol{\beta}$ not depending on $\boldsymbol{\beta}$ (recall $\dot{\boldsymbol{\ell}}_{n_j}^{(j)}(\boldsymbol{\beta})/n_j \xrightarrow{p} \dot{\mathbf{d}}_y^{(j)}(\boldsymbol{\beta})$). Hence we have

$$
\begin{aligned}
& n_{\text{pool}}^{1/2} \Big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}^* \Big) = n_{\text{pool}}^{1/2} \Big(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} \Big)^{-1} \sum_{j=1}^{D} \Big[\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(j)} \big(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(j)} - \boldsymbol{\beta}^* \big) \Big] \\
=& - \Big(\frac{1}{n_{\text{pool}}} \sum_{k=1}^{D} \ddot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}^*) \Big)^{-1} \frac{1}{n_{\text{pool}}^{1/2}} \sum_{k=1}^{D} \dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}^*) + o_p(1) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{A}^{\text{pool}}(\boldsymbol{\beta}^*)^{-1} \mathbf{B}^{\text{pool}}(\boldsymbol{\beta}^*) \mathbf{A}^{\text{pool}}(\boldsymbol{\beta}^*)^{-1} \big)
\end{aligned}
$$

following Eq. (26) in Appendix E.1, $\mathbf{A}^{\text{pool}}(\boldsymbol{\beta}^*) = \sum_{k=1}^{D} p_k \mathcal{I}^{(k)}(\boldsymbol{\beta}^*)$, and $\mathbf{B}^{\text{pool}}(\boldsymbol{\beta}^*) = \sum_{k=1}^{D} p_k \mathbb{E}\big[\dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}^*) \dot{\boldsymbol{\ell}}_{n_k}^{(k)}(\boldsymbol{\beta}^*)^\top \big]$ under Assumption 2. We then complete the proof of Proposition 5. □

E.4 Proof of Results for Federated IPW-MLE in Section 4.2

E.4.1 Proof of Lemma 1

Proof of Lemma 1. Suppose the propensity model is the same across all data sets. Let us first show the asymptotic distribution for $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ when the propensity is estimated. We parameterize the propensity score as $\text{pr}(W_i|\mathbf{X}_i) = e(\mathbf{X}_i, \boldsymbol{\gamma})$, and the corresponding maximum likelihood estimator is denoted as $\hat{\boldsymbol{\gamma}}$. Furthermore, we denote the likelihood of W_i given $\mathbf{X}_i$ and $\boldsymbol{\gamma}$ as $f(W_i|\mathbf{X}_i, \boldsymbol{\gamma})$, and then we have $e(\mathbf{X}_i, \boldsymbol{\gamma}) = f(W_i = 1|\mathbf{X}_i, \boldsymbol{\gamma})$.

It is possible for $e(\mathbf{x}, \boldsymbol{\gamma})$ to be misspecified (Condition 7 is violated). In this case, under regularity

conditions in White (1982), $\hat{\boldsymbol{\gamma}}_{\text{mle}}$ is consistent and asymptotically normal:

$$\sqrt{n}(\hat{\boldsymbol{\gamma}}_{\text{mle}} - \boldsymbol{\gamma}^*) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{V}_{\boldsymbol{\gamma}}\big), \tag{32}$$

where $\boldsymbol{\gamma}^*$ minimizes the Kullback-Leibler Information Criterion between the true model and the parameterized model $e(\mathbf{X}_i, \boldsymbol{\gamma}^*)$, and $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}_{\boldsymbol{\gamma}^*}^{-1}\mathbf{B}_{\boldsymbol{\gamma}^*}\mathbf{A}_{\boldsymbol{\gamma}^*}^{-1}$. $\mathbf{A}_{\boldsymbol{\gamma}^*}$ is $\mathbf{A}_{\boldsymbol{\gamma}}$ evaluated at $\boldsymbol{\gamma}^*$ with the definition of $\mathbf{A}_{\boldsymbol{\gamma}}$ provided in Table 1, and likewise for $\mathbf{B}_{\boldsymbol{\gamma}^*}$.

Note that $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ satisfies the first order condition of the objective function (4). With probability approaching one, we have the mean value expansion of the first order condition (or score) at $\boldsymbol{\beta}_0$ of:

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varpi_{i,\hat{e}}\mathbf{g}(\mathbf{X}_i, W_i, \boldsymbol{\beta}_0) + \left(\frac{1}{n}\sum_{i=1}^{n} \varpi_{i,\hat{e}}\ddot{\mathbf{H}}(\mathbf{X}_i, W_i, \tilde{\boldsymbol{\beta}})\right)\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\big),$$

where $\mathbf{g}_i \coloneqq \mathbf{g}(\mathbf{X}_i, W_i, \boldsymbol{\beta}_0) = \frac{\partial}{\partial \boldsymbol{\beta}} \log f(Y_i|\mathbf{X}_i, W_i, \boldsymbol{\beta}_0)$, $\ddot{\mathbf{H}}(\mathbf{X}_i, W_i, \tilde{\boldsymbol{\beta}}) = \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} \log f(Y_i|\mathbf{X}_i, W_i, \tilde{\boldsymbol{\beta}})$ with $\tilde{\boldsymbol{\beta}}$ lying between $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ and $\boldsymbol{\beta}_0$, and $\varpi_{i,\hat{e}} = \frac{W_i}{\hat{e}(\mathbf{X}_i)} + \frac{1-W_i}{1-\hat{e}(\mathbf{X}_i)}$ for ATE weighting or $\varpi_{i,\hat{e}} = W_i + \frac{\hat{e}(\mathbf{X}_i)}{1-\hat{e}(\mathbf{X}_i)}(1 - W_i)$ for ATT weighting.

By the uniform weak law of large numbers, we have

$$\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\big) = -\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varpi_{i,\hat{e}}\mathbf{g}_i\right) + o_p(1),$$

where $\mathbf{A}_{\boldsymbol{\beta}_0,\varpi} = \frac{1}{n}\sum_{i=1}^{n} \varpi_{i,\hat{e}}\ddot{\mathbf{H}}(\mathbf{X}_i, W_i, \boldsymbol{\beta}_0)$. The next step is to use the mean value expansion on $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varpi_{i,\hat{e}}\mathbf{g}_i$ at $\boldsymbol{\gamma}^*$; we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varpi_{i,\hat{e}}\mathbf{g}_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \underbrace{\varpi_{i,e_{\boldsymbol{\gamma}^*}}\mathbf{g}_i}_{\coloneqq \mathbf{k}_i} + \mathbb{E}\Big[\mathbf{g}_i\Big(\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\Big)^\top\Big]\sqrt{n}(\hat{\boldsymbol{\gamma}}_{\text{mle}} - \boldsymbol{\gamma}^*) + o_p(1),$$

where $\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}$ is the first order derivative of $\varpi_{i,e_{\boldsymbol{\gamma}}}$ with respect to $\boldsymbol{\gamma}$ evaluated at $\boldsymbol{\gamma}^*$. In order to show the asympototic distribution of $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$, we need to show the asymptotic distribution of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varpi_{i,\hat{e}}\mathbf{g}_i$. We analyze the leading terms in the above equation one by one.

Let us first consider the ATE weighting. In this case, $\varpi_{i,e} = \frac{W_i}{e(\mathbf{X}_i,\boldsymbol{\gamma})} + \frac{1-W_i}{1-e(\mathbf{X}_i,\boldsymbol{\gamma})}$ and

$$\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} = -\frac{W_i}{(e_i^*)^2}\frac{\partial e(\mathbf{X}_i, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}} - \frac{1 - W_i}{(1 - e_i^*)^2}\frac{\partial(1 - e(\mathbf{X}_i, \boldsymbol{\gamma}^*))}{\partial \boldsymbol{\gamma}},$$

where $e_i^* = e(\mathbf{X}_i, \boldsymbol{\gamma}^*)$. Under Asssumption 1 and the asymptotic distribution (26) in Appendix E.1, we have

$$\sqrt{n}\big(\hat{\boldsymbol{\gamma}}_{\text{mle}} - \boldsymbol{\gamma}^*\big) = \mathbf{A}_{\boldsymbol{\gamma}^*}^{-1} \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{d}_i + o_p(1),$$

where $\mathbf{A}_{\boldsymbol{\gamma}^*}$ is $\mathbf{A}_{\boldsymbol{\gamma}}$ evaluated at $\boldsymbol{\gamma}^*$, the definition of $\mathbf{A}_{\boldsymbol{\gamma}}$ can be found in Table 1, and $\mathbf{d}_i$ is defined

as

$$\mathbf{d}_i = \frac{W_i}{e_i^*}\frac{\partial e(\mathbf{X}_i,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}} - \frac{1-W_i}{1-e_i^*}\frac{\partial e(\mathbf{X}_i,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}},$$

which is the first order derviative (or score) of the binary response (treatment variable W_i) evaluated at $\boldsymbol{\gamma}^*$. If $e(\mathbf{X}_i,\boldsymbol{\gamma})$ is correctly specified, we have $\mathbf{A}_{\boldsymbol{\gamma}^*} = \mathbb{E}[\mathbf{d}_i\mathbf{d}_i^\top]$. Using $W_i(1-W_i)=0$, we have $W_i\Big(\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\Big) = -\frac{W_i}{e_i^*}\mathbf{d}_i$ and $(1-W_i)\Big(\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\Big) = -\frac{1-W_i}{1-e_i^*}\mathbf{d}_i$. Therefore,

$$\mathbb{E}\Big[\mathbf{g}_i\Big(\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\Big)^\top\Big] = -\mathbb{E}\Big[\underbrace{\Big(\frac{W_i}{e_i^*}+\frac{1-W_i}{1-e_i^*}\Big)\mathbf{g}_i}_{\mathbf{k}_i}\mathbf{d}_i^\top\Big].$$

Collecting terms together, we have shown

$$\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\big) = -\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{k}_i - \mathbb{E}\big[\mathbf{k}_i\mathbf{d}_i^\top\big]\mathbf{A}_{\boldsymbol{\gamma}^*}^{-1}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{d}_i\right) + o_p(1). \tag{33}$$

Since the standard unconfoundedness assumption holds (stated in Section 2.1), the randomness of $\mathbf{k}_i$ comes from the residual in Y_i, and the randomness of $\mathbf{d}_i$ comes from the residual in W_i, and we have $\mathbf{k}_i$ uncorrelated with $\mathbf{d}_j$ for any i and j (including the case where i and j are the same). In addition, observations are i.i.d., $\mathbf{k}_i$ is uncorrelated with $\mathbf{k}_j$, and $\mathbf{d}_i$ is uncorrelated with $\mathbf{d}_j$ for $i \neq j$. Then, we have

$$\mathbf{V}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}^\dagger = \mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{-1}\Big(\underbrace{\mathbb{E}\big[\mathbf{k}_i\mathbf{k}_i^\top\big]}_{\mathbf{D}_{\boldsymbol{\beta}_0,\varpi}} - \underbrace{\mathbb{E}\big[\mathbf{k}_i\mathbf{d}_i^\top\big]}_{\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}}\mathbf{V}_{\boldsymbol{\gamma}}\underbrace{\mathbb{E}\big[\mathbf{d}_i\mathbf{k}_i^\top\big]}_{\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}^\top}\Big)\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{-1},$$

where $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}_{\boldsymbol{\gamma}^*}^{-1}\mathbf{B}_{\boldsymbol{\gamma}^*}\mathbf{A}_{\boldsymbol{\gamma}^*}^{-1}$. If $e(\mathbf{X}_i,\boldsymbol{\gamma})$ is correctly specified, we have $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbb{E}[\mathbf{d}_i\mathbf{d}_i^\top]^{-1} = \mathbf{A}_{\boldsymbol{\gamma}_0}^{-1}$.

If we use the true propensity score, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \varpi_{i,\hat{e}}\mathbf{g}_i = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varpi_{i,e}\mathbf{g}_i + o_p(1)$$

and

$$\mathbf{V}_{\boldsymbol{\beta},\text{ipw-mle},e}^\dagger = \mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{-1}\underbrace{\mathbb{E}\big[\mathbf{k}_i\mathbf{k}_i^\top\big]}_{\mathbf{D}_{\boldsymbol{\beta}_0,\varpi}}\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{-1}.$$

Next, let us consider the ATT weighting. In this case, $\varpi_{i,e} = W_i + \frac{e(\mathbf{X}_i,\boldsymbol{\gamma})}{1-e(\mathbf{X}_i,\boldsymbol{\gamma})}(1-W_i)$ and

$$\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} = \frac{1-W_i}{(1-e_i^*)^2}\frac{\partial e(\mathbf{X}_i,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}}.$$

Using $W_i(1-W_i)=0$, we have $\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*} = -\frac{1-W_i}{1-e_i^*}\mathbf{d}_i$. Therefore,

$$\mathbb{E}\Big[\mathbf{g}_i\Big(\frac{\partial \varpi_{i,e_{\boldsymbol{\gamma}}}}{\partial \boldsymbol{\gamma}}\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^*}\Big)^\top\Big] = -\mathbb{E}\Big[\underbrace{\frac{1-W_i}{1-e_i^*}\mathbf{g}_i}_{\mathbf{h}_i}\mathbf{d}_i^\top\Big].$$

If the propensity score is estimated, then $\mathbf{V}^\dagger_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ takes the form of

$$\begin{aligned}&\mathbf{V}^\dagger_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}\\ =&\mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi}\Big(\underbrace{\mathbb{E}\big[\mathbf{k}_i\mathbf{k}_i^\top\big]}_{\mathbf{D}_{\boldsymbol{\beta}_0,\varpi}} - \underbrace{\mathbb{E}\big[\mathbf{h}_i\mathbf{d}_i^\top\big]}_{\mathbf{C}_{\boldsymbol{\beta}_0,\varpi,1}}\mathbf{V}_{\boldsymbol{\gamma}}\underbrace{\mathbb{E}\big[\mathbf{d}_i\mathbf{k}_i^\top\big]}_{\mathbf{C}^\top_{\boldsymbol{\beta}_0,\varpi,2}} - \underbrace{\mathbb{E}\big[\mathbf{k}_i\mathbf{d}_i^\top\big]}_{\mathbf{C}_{\boldsymbol{\beta}_0,\varpi,2}}\mathbf{V}_{\boldsymbol{\gamma}}\underbrace{\mathbb{E}\big[\mathbf{d}_i\mathbf{h}_i^\top\big]}_{\mathbf{C}^\top_{\boldsymbol{\beta}_0,\varpi,1}} + \underbrace{\mathbb{E}\big[\mathbf{h}_i\mathbf{d}_i^\top\big]}_{\mathbf{C}_{\boldsymbol{\beta}_0,\varpi,2}}\mathbf{V}_{\boldsymbol{\gamma}}\underbrace{\mathbb{E}\big[\mathbf{d}_i\mathbf{h}_i^\top\big]}_{\mathbf{C}^\top_{\boldsymbol{\beta}_0,\varpi,2}}\Big)\mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi},\end{aligned}$$

where $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbf{A}^{-1}_{\boldsymbol{\gamma}^*}\mathbf{B}_{\boldsymbol{\gamma}^*}\mathbf{A}^{-1}_{\boldsymbol{\gamma}^*}$. If $e(\mathbf{X}_i,\boldsymbol{\gamma})$ is correctly specified, we have $\mathbf{V}_{\boldsymbol{\gamma}} = \mathbb{E}[\mathbf{d}_i\mathbf{d}_i^\top]^{-1} = \mathbf{A}^{-1}_{\boldsymbol{\gamma}_0}$. If we use the true propensity score, then

$$\mathbf{V}^\dagger_{\boldsymbol{\beta},\text{ipw-mle},e} = \mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi}\underbrace{\mathbb{E}\big[\mathbf{k}_i\mathbf{k}_i^\top\big]}_{\mathbf{D}_{\boldsymbol{\beta}_0,\varpi}}\mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi}.$$

□

E.4.2 Proof of Theorem 2

Proof of Theorem 2 . In this proof, we show the results for the federated estimators where the estimated propensity is used. If the true propensity is used (Condition 4), we can follow the same procedure to prove the results for this case. Our proof of Theorem 2 consists of showing the following four equations

1. $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0,\mathbf{I}_d)$
2. $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0,\mathbf{I}_d)$
3. $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0,\mathbf{I}_d)$
4. $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0,\mathbf{I}_d)$.

Similar to the proof of Theorem 1, we need to consider two cases in this proof. The first case is the information matrix $\mathcal{I}^{(k)}(\boldsymbol{\beta})$ being the same for all data sets, in which case Assumption 2 is not required. The second case is the information matrix varying across data sets, in which case we require Assumption 2.

The first step is to show $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0,\mathbf{I}_d)$. From Lemma 1, for

the combined data (that can be viewed as a single data set), we have

$$\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}} \xrightarrow{p} \boldsymbol{\beta}_0,$$
$$\sqrt{n_k}\big(\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\big) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}\big),$$

where $\mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}$ is the asymptotic variance (see Lemma 1 for its expression). From the law of large numbers and the consistency of $\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}}$, we have $\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ be a consistent estimator of $\mathbf{V}^{\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$. Hence, by Slutsky's theorem, we have

$$n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d).$$

The second step is to show the second equation (i.e., $n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$) for the case where the case where $\mathbf{V}^{(k),\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}$ is the same for all data sets (and we do not require Assumption 2 to hold).

For this case, we drop superscript k for notation simplicity. In order to show the second equation, we need to additionally show the consistency of $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ given what we have in the first step. To show the consistency of $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$, we start with showing the consistency of $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$ and $\hat{\boldsymbol{\gamma}}^{\text{fed}}_{\text{mle}}$. We can follow the same procedure as the proof of $\left\|\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{mle}} - \boldsymbol{\beta}_0\right\|_2 = o_p(1)$ in Inequality (30) (in the proof of Theorem 1) to show the consistency of $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$ and $\hat{\boldsymbol{\gamma}}^{\text{fed}}_{\text{mle}}$.

In more detail, for $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$ (recall we use Hessian weighting to pool $\hat{\boldsymbol{\beta}}^{(k)}_{\text{ipw-mle}}$, denoting the Hessian on data set k as $\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta},\text{ipw-mle}}$ and $\hat{p}_{n,j} = \frac{n_j}{\sum_{k=1}^D n_k}$),

$$\begin{aligned}
\left\|\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\right\|_2 &= \left\|\Big(\sum_{k=1}^D \hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta},\text{ipw-mle}}\Big)^{-1}\Big(\sum_{k=1}^D \hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta},\text{ipw-mle}}\big(\hat{\boldsymbol{\beta}}^{(k)}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\big)\Big)\right\|_2 \\
&\le \sum_{j=1}^D \hat{p}_{n,j} \cdot \underbrace{\left\|\Big(\sum_{k=1}^D \hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta},\text{ipw-mle}}\Big)^{-1}\hat{\mathbf{H}}^{(j)}_{\boldsymbol{\beta},\text{ipw-mle}} \cdot \frac{1}{\hat{p}_{n,j}} \cdot \Big(\hat{\boldsymbol{\beta}}^{(j)}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\Big)\right\|_2}_{o_p(1)} = o_p(1),
\end{aligned}$$

where we use the property that $\big(\sum_{k=1}^D \hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta},\text{ipw-mle}}\big)^{-1}\hat{\mathbf{H}}^{(j)}_{\boldsymbol{\beta},\text{ipw-mle}}\frac{1}{\hat{p}_{n,j}} \xrightarrow{p} \mathbf{I}_d$ (which can be shown in the same procedure as Eq. (29), where we additionally use the consistency of $\hat{e}^{(k)}$). Therefore we finish the proof of the consistency of $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$.

Next we show the consistency of $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$. Recall from Table 3 that in the estimation of $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$, we use $\hat{\mathbf{A}}^{(k)}_{\boldsymbol{\beta},\varpi}, \hat{\mathbf{C}}^{(k)}_{\boldsymbol{\beta},\varpi}, \hat{\mathbf{D}}^{(k)}_{\boldsymbol{\beta},\varpi}, \hat{\mathbf{A}}^{(k)}_{\boldsymbol{\gamma}}$, and $\hat{\mathbf{B}}^{(k)}_{\boldsymbol{\gamma}}$ (for ATT weighting, replace $\hat{\mathbf{C}}^{(k)}_{\boldsymbol{\beta},\varpi}$ by $\hat{\mathbf{C}}^{(k)}_{\boldsymbol{\beta},\varpi,1}, \hat{\mathbf{C}}^{(k)}_{\boldsymbol{\beta},\varpi,2}$) which are estimated using $\hat{\boldsymbol{\gamma}}^{\text{fed}}$ and $\hat{\boldsymbol{\beta}}^{\text{fed}}$. By the uniform weak law of large numbers, all these quantities are consistent. Using exactly the same proof that showed $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} \xrightarrow{p} \boldsymbol{\beta}_0$, we can show the consistency of $\hat{\mathbf{A}}^{\text{fed}}_{\boldsymbol{\beta},\varpi}, \hat{\mathbf{C}}^{\text{fed}}_{\boldsymbol{\beta},\varpi}, \hat{\mathbf{D}}^{\text{fed}}_{\boldsymbol{\beta},\varpi}, \hat{\mathbf{A}}^{\text{fed}}_{\boldsymbol{\gamma}}$, and $\hat{\mathbf{B}}^{\text{fed}}_{\boldsymbol{\gamma}}$ (for ATT weighting, replace $\hat{\mathbf{C}}^{\text{fed}}_{\boldsymbol{\beta},\varpi}$

by $\hat{\mathbf{C}}^{\text{fed}}_{\boldsymbol{\beta},\varpi,1}, \hat{\mathbf{C}}^{\text{fed}}_{\boldsymbol{\beta},\varpi,2}$). Then, the consistency of $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ can be shown:

$$\begin{aligned}\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}} =& \big(\hat{\mathbf{A}}^{\text{fed}}_{\boldsymbol{\beta},\varpi}\big)^{-1}\big(\hat{\mathbf{D}}^{\text{fed}}_{\boldsymbol{\beta},\varpi} - \hat{\mathbf{M}}^{\text{fed}}_{\boldsymbol{\beta},\varpi,\boldsymbol{\gamma}}\big)\big(\hat{\mathbf{A}}^{\text{fed}}_{\boldsymbol{\beta},\varpi}\big)^{-1} \\ &\xrightarrow{p} \mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi}\big(\mathbf{D}_{\boldsymbol{\beta}_0,\varpi} - \mathbf{M}_{\boldsymbol{\beta}_0,\varpi,\boldsymbol{\gamma}}\big)\mathbf{A}^{-1}_{\boldsymbol{\beta}_0,\varpi} = \mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}},\end{aligned} \tag{34}$$

where $\hat{\mathbf{M}}^{\text{fed}}_{\boldsymbol{\beta},\varpi,\boldsymbol{\gamma}}$ is a smooth function of $\hat{\mathbf{C}}^{\text{fed}}_{\boldsymbol{\beta},\varpi}, \hat{\mathbf{A}}^{\text{fed}}_{\boldsymbol{\gamma}}$ and $\hat{\mathbf{B}}^{\text{fed}}_{\boldsymbol{\gamma}}$ for ATE weighting, and $\hat{\mathbf{M}}^{\text{fed}}_{\boldsymbol{\beta},\varpi,\boldsymbol{\gamma}}$ is a smooth function of $\hat{\mathbf{C}}^{\text{fed}}_{\boldsymbol{\beta},\varpi,1}, \hat{\mathbf{C}}^{\text{fed}}_{\boldsymbol{\beta},\varpi,2}, \hat{\mathbf{A}}^{\text{fed}}_{\boldsymbol{\gamma}}$, and $\hat{\mathbf{B}}^{\text{fed}}_{\boldsymbol{\gamma}}$ for ATT weighting.

Given the consistency of $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$, we have recovered the second equation:

$$n^{1/2}_{\text{pool}}(\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{pool}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$$

The third step is to show the third and fourth equations together for the case where the case where $\mathbf{V}^{(k),\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}$ is the same for all data sets (i.e., $n^{1/2}_{\text{pool}}(\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$ and $n^{1/2}_{\text{pool}}(\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}})^{-1/2}(\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_d)$). Given the consistency of $\hat{\mathbf{V}}^{\text{pool},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ and $\hat{\mathbf{V}}^{\text{fed},\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$ (from the proofs of the first and second equations), if we can show $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$ converges to $\boldsymbol{\beta}_0$ in an asymptotic normal distribution with the convergence rate $n^{1/2}_{\text{pool}}$ and asymptotic variance with the asymptotic variance $\mathbf{V}^{\dagger}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}$, then by Slutsky's theorem, we obtain the third and fourth equations.

Since observations between data sets are asymptotically independent, we have that $\Big(n^{1/2}_1(\hat{\boldsymbol{\beta}}^{(1)}_{\text{ipw-mle}} - \boldsymbol{\beta}_0), n^{1/2}_2(\hat{\boldsymbol{\beta}}^{(2)}_{\text{ipw-mle}} - \boldsymbol{\beta}_0), \cdots, n^{1/2}_D(\hat{\boldsymbol{\beta}}^{D}_{\text{ipw-mle}} - \boldsymbol{\beta}_0)\Big)$ converges jointly to a normal distribution, for any $j \neq k$, $n^{1/2}_j(\hat{\boldsymbol{\beta}}^{(j)}_{\text{ipw-mle}} - \boldsymbol{\beta}_0)$ and $n^{1/2}_k(\hat{\boldsymbol{\beta}}^{(k)}_{\text{ipw-mle}} - \boldsymbol{\beta}_0)$ are independent, and

$$n^{1/2}_{\text{pool}}\Big(\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0\Big) = \sum_{j=1}^{D} \hat{p}^{1/2}_{n,j}\Big[\underbrace{\big(\sum_{k=1}^{D}\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta},\text{ipw-mle}}\big)^{-1}\hat{\mathbf{H}}^{(j)}_{\boldsymbol{\beta},\text{ipw-mle}}\frac{1}{\hat{p}_{n,j}} \cdot n^{1/2}_j(\hat{\boldsymbol{\beta}}^{(j)}_{\text{ipw-mle}} - \boldsymbol{\beta}_0)}_{:=\boldsymbol{\xi}^{(j)}_{n_j} \xrightarrow{d} \mathcal{N}(0, \mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}) \text{ from } \left(\sum_{k=1}^{D}\hat{\mathbf{H}}^{(k)}_{\boldsymbol{\beta},\text{ipw-mle}}\right)^{-1}\hat{\mathbf{H}}^{(j)}_{\boldsymbol{\beta},\text{ipw-mle}}\frac{1}{\hat{p}_{n,j}} \xrightarrow{p} \mathbf{I}_d \text{ and Slutsky's theorem}}\Big].$$

We can use the almost sure representation argument, as in the proof of Theorem 1, to show

$$\sum_{j=1}^{D}\hat{p}^{1/2}_{n,j}\boldsymbol{\xi}^{j}_{n_j} \xrightarrow{d} \mathcal{N}\Big(0, \sum_{j=1}^{D} p_{n,j}\mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}\Big) \overset{d}{=} \mathcal{N}\Big(0, \mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}\Big)$$

and the asymptotic distribution of $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$

$$n^{1/2}_{\text{pool}}(\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\Big(0, \mathbf{V}^{\dagger}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}\Big).$$

By Slutsky's theorem, we finish showing the third and fourth equations,

$$n_{\text{pool}}^{1/2}\big(\hat{\mathbf{V}}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}^{\text{fed},\dagger}\big)^{-1/2}\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}}-\boldsymbol{\beta}_0\big)\xrightarrow{d}\mathcal{N}\big(0,\mathbf{I}_d\big)$$
$$n_{\text{pool}}^{1/2}\big(\hat{\mathbf{V}}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}^{\text{pool},\dagger}\big)^{-1/2}\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}}-\boldsymbol{\beta}_0\big)\xrightarrow{d}\mathcal{N}\big(0,\mathbf{I}_d\big).$$

The last step is to show the second to fourth equations for the case where $\mathbf{V}_{\boldsymbol{\beta}_0,\text{ipw-mle},\hat{e}}^{(k),\dagger}$ differs across data sets under Assumption 2. Based on what we have from the first case, we only need to additionally show that $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}}$ and $\hat{\mathbf{V}}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}^{\text{fed},\dagger}$ are consistent and $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}}$ is asymptotically normal with variance $\mathbf{V}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}^{\dagger}$ even when $\mathbf{V}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}^{(k),\dagger}$ differs across data sets.

Let us start with the consistency of $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}}$. Recall from our federation procedure of the IPW-MLE estimator that we first estimate the propensity model on the combined data and use this federated propensity model to estimate $\boldsymbol{\beta}_{\text{ipw-mle}}^{(k)}$ on each data set. Then, for the ATE weighting, the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{(k)}$ satisfies (ATT weighting can be shown analogously with a similar equation):

$$\begin{aligned}n_k^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{(k)}-\boldsymbol{\beta}_0\big)&=-\big(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{(k)}\big)^{-1}\left(\frac{1}{n_j^{1/2}}\sum_{i=1}^{n_k}\mathbf{k}_i-\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}^{(k)}\cdot\big(\mathbf{A}_{\boldsymbol{\gamma}^*}^{\text{pool}}\big)^{-1}\cdot\hat{p}_{n,k}^{1/2}\cdot\frac{1}{n_{\text{pool}}^{1/2}}\sum_{i=1}^{n_{\text{pool}}}\mathbf{d}_i\right)+o_p(1)\\&\xrightarrow{d}\mathcal{N}\left(0,\big(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{(k)}\big)^{-1}\Big(\mathbf{D}_{\boldsymbol{\beta}_0,\varpi}^{(k)}-\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}^{(k)}\cdot p_k\mathbf{V}_{\boldsymbol{\gamma}}^{\text{pool}}\cdot\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}^{(k)}\Big)\big(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{(k)}\big)^{-1}\right)\end{aligned}$$

given Assumption 2, where the definitions of $\mathbf{k}_i$ and $\mathbf{d}_i$ can be found in the proof of Lemma 1. Note that we have $\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(k)}/n_k\xrightarrow{p}\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{(k)}$. Since $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{(k)}$ is consistent, we have $\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(k)}/n_{\text{pool}}\xrightarrow{p}\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}}$, and therefore $\big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(k)}\big)^{-1}\cdot\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(j)}\cdot\frac{1}{\hat{p}_{n,j}}\xrightarrow{p}(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}})^{-1}\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{(k)}$. Given the assumption $\Big\|(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}})^{-1}\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{(k)}\Big\|_2\le M$, then $\big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(k)}\big)^{-1}\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(j)}\cdot\frac{1}{\hat{p}_{n,j}}\cdot\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{(k)}-\boldsymbol{\beta}_0\big)=o_p(1)$ continues to hold, and therefore $\Big\|\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}}-\boldsymbol{\beta}_0\Big\|_2=o_p(1)$ (where $\hat{\boldsymbol{\gamma}}_{\text{mle}}^{\text{fed}}\xrightarrow{p}\boldsymbol{\gamma}_0$ can be shown using exactly the same proof).

Lastly, we show the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}}$. Using $\big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(k)}\big)^{-1}\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(j)}\cdot\frac{1}{\hat{p}_{n,j}}\xrightarrow{p}(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}})^{-1}\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{(k)}$, we have the following for ATE weighting (with similar arithmetic for ATT weighting):

$$\begin{aligned}&n_{\text{pool}}^{1/2}\Big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{\text{fed}}-\boldsymbol{\beta}_0\Big)=-\sum_{j=1}^{D}\hat{p}_{n,j}^{1/2}\Big[\big(\sum_{k=1}^{D}\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(k)}\big)^{-1}\hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ipw-mle}}^{(j)}\cdot\frac{1}{\hat{p}_{n,j}}\cdot n_j^{1/2}\big(\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}^{(j)}-\boldsymbol{\beta}_0\big)\Big]\\
&=-\big(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}}\big)^{-1}\sum_{j=1}^{D}\hat{p}_{n,j}^{1/2}\left(\frac{1}{n_j^{1/2}}\sum_{i=1}^{n_j}\mathbf{k}_i-\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}^{(j)}\cdot\big(\mathbf{A}_{\boldsymbol{\gamma}^*}^{\text{pool}}\big)^{-1}\cdot\hat{p}_{n,j}^{1/2}\cdot\frac{1}{n_{\text{pool}}^{1/2}}\sum_{i=1}^{n_{\text{pool}}}\mathbf{d}_i\right)+o_p(1)\\
&=-\frac{\big(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}}\big)^{-1}}{n_{\text{pool}}^{1/2}}\left(\sum_{i=1}^{n_{\text{pool}}}\mathbf{k}_i-\left(\sum_{j=1}^{D}\frac{n_j}{n_{\text{pool}}}\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}^{(j)}\right)\cdot\big(\mathbf{A}_{\boldsymbol{\gamma}^*}^{\text{pool}}\big)^{-1}\cdot\sum_{i=1}^{n_{\text{pool}}}\mathbf{d}_i\right)+o_p(1)\\
&\xrightarrow{d}\mathcal{N}\left(0,\big(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}}\big)^{-1}\Big(\mathbf{D}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}}-\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}}\cdot\mathbf{V}_{\boldsymbol{\gamma}}^{\text{pool}}\cdot\mathbf{C}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}}\Big)\big(\mathbf{A}_{\boldsymbol{\beta}_0,\varpi}^{\text{pool}}\big)^{-1}\right)\equiv\mathcal{N}\Big(0,\mathbf{V}_{\boldsymbol{\beta},\text{ipw-mle},\hat{e}}^{\dagger}\Big).\end{aligned}$$

We have hence shown the asymptotic distribution of $\hat{\boldsymbol{\beta}}^{\text{fed}}_{\text{ipw-mle}}$, which completes the proof in the second case. □

E.4.3 Proof of Proposition 4

Proof of Proposition 4. We can follow the same approach as Proposition 3 and Theorem 2 to show this proposition; details are therefore omitted and available upon request. □

E.5 Proof of Results for Federated AIPW in Section 4.3

Proof of Theorem 3. In order to prove Theorem 3, let us first review some properties of $\hat{\tau}^{\text{aipw}}$ estimated from a single data set. If either the propensity or outcome model is correctly specified, $\hat{\tau}_{\text{aipw}}$ is asymptotically linear (Tsiatis and Davidian, 2007),

$$\sqrt{n}\big(\hat{\tau}_{\text{aipw}} - \tau_0\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(\mathbf{X}_i, W_i, Y_i) + o_p(1) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{V}_\tau\big), \tag{35}$$

where $\phi(\mathbf{x}, w, y)$ is an influence function that satisfies $\mathbb{E}[\phi(\mathbf{x}, w, y)] = 0$ and $\mathbf{V}_\tau = \mathbb{E}[\phi(\mathbf{x}, w, y)^2] < \infty$. Suppose the score function of $s(\mathbf{X}_i, W_i, Y_i)$ can be parameterized by $\boldsymbol{\theta}$, with the true value being $\boldsymbol{\theta}_0$; then, the treatment effect τ_0 can also be parameterized, i.e., $\tau_0 = \tau(\boldsymbol{\theta}_0)$, and τ_0 is differentiable in $\boldsymbol{\theta}$. From Newey (1994), $\phi(\mathbf{X}_i, W_i, Y_i)$ as a valid influence function connects τ_0 and $s(\mathbf{X}_i, W_i, Y_i|\boldsymbol{\theta})$ via

$$\frac{\partial \tau(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \mathbb{E}\big[\phi(\mathbf{X}_i, W_i, Y_i) s(\mathbf{X}_i, W_i, Y_i|\boldsymbol{\theta}_0)\big]. \tag{36}$$

Now we are ready to show Theorem 3. We aim to find a valid influence function that satisfies the equality (36) on the combined data, and then we can use this valid influence function to provide the asymptotic distribution of $\hat{\tau}^{\text{pool}}_{\text{aipw}}$ and $\hat{\tau}^{\text{fed}}_{\text{aipw}}$. The population treatment effect and score function on the combined data set satisfy the following (recall that $p_j = \lim \frac{n_j}{n_{\text{pool}}}$ as defined in Assumption 2):

$$\begin{aligned}
\tau_0 &= \sum_{j=1}^{D} p_j \tau_0^{(j)} \\
s^{\text{pool}}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}|\boldsymbol{\theta}_0) &= \sum_{j=1}^{D} \mathbb{1}(k=j) s^{(j)}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}|\boldsymbol{\theta}_0^{(j)}).
\end{aligned}$$

Let a candidate influence function on the combined data set be

$$\phi^{\text{pool}}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}) = \sum_{j=1}^{D} \mathbb{1}(k=j) \phi^{(j)}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}).$$

This candidate influence function satisfies $\mathbb{E}[\phi^{\text{pool}}(\mathbf{x}, w, y)] = 0$, $\mathbb{E}[\phi^{\text{pool}}(\mathbf{x}, w, y)^2] < \infty$,

$$\phi^{\text{pool}}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}) = \phi^{(k)}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}), \tag{37}$$

and

$$\begin{aligned}\frac{\partial \tau(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} =& \sum_{j=1}^{D} p_j \frac{\partial \tau(\boldsymbol{\theta}_0^{(j)})}{\partial \boldsymbol{\theta}^{(j)}} = \sum_{j=1}^{D} p_j \mathbb{E}\big[\phi^{(j)}(\mathbf{X}_i^{(j)}, W_i^{(j)}, Y_i^{(j)}) s(\mathbf{X}_i^{(j)}, W_i^{(j)}, Y_i^{(j)} | \boldsymbol{\theta}_0^{(j)})\big] \\ =& \mathbb{E}\big[\phi^{\text{pool}}(\mathbf{X}_i, W_i, Y_i) s^{\text{pool}}(\mathbf{X}_i, W_i, Y_i | \boldsymbol{\theta}_0)\big],\end{aligned}$$

i.e., equality (36) holds for $\phi^{\text{pool}}(\mathbf{X}_i, W_i, Y_i)$, and therefore, $\phi^{\text{pool}}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)})$ is a valid influence function. Based on this influence function, we have

$$n_{\text{pool}}^{1/2}\big(\hat{\tau}_{\text{aipw}}^{\text{pool}} - \tau_0\big) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{V}_\tau^{\text{pool}}\big)$$

where the asymptotic variance $\mathbf{V}_\tau^{\text{pool}}$ satisfies

$$\mathbf{V}_\tau^{\text{pool}} = \mathbb{E}\big[\phi^{\text{pool}}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)})^2\big] = \sum_{j=1}^{D} p_j \mathbb{E}\big[\phi^{(j)}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)})^2\big] = \sum_{j=1}^{D} p_j \mathbf{V}_\tau^{(k)}$$

using the property that $\mathbb{1}(k = j) \cdot \mathbb{1}(k = l) = 0$ for $j \neq l$, where $\mathbf{V}_\tau^{(k)}$ is the asymptotic variance on data set k.

$\hat{\mathbf{V}}_\tau^{\text{pool}}$ is consistent from Lemma 2 and the definition of $\hat{\mathbf{V}}_\tau^{\text{pool}}$, and from Slutsky's theorem, we have

$$n_{\text{pool}}^{1/2}(\hat{\mathbf{V}}_\tau^{\text{pool}})^{-1/2}(\hat{\tau}_{\text{aipw}}^{\text{pool}} - \tau_0) \xrightarrow{d} \mathcal{N}(0, 1).$$

For the case where $\phi(\mathbf{X}_i, W_i, Y_i)$ varies with the data set, the federated treatment effect $\hat{\tau}_{\text{aipw}}^{\text{fed}}$ from sample size weighting in Section 3.3.2 satisfies

$$\begin{aligned}n_{\text{pool}}^{1/2}(\hat{\tau}_{\text{aipw}}^{\text{fed}} - \tau_0) =& n_{\text{pool}}^{1/2} \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \cdot \frac{1}{n_k} \sum_{i=1}^{n_k} \phi^{(k)}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}) + o_p(1) \\ =& \frac{1}{n_{\text{pool}}^{1/2}} \sum_{k=1}^{D} \sum_{i=1}^{n_k} \phi^{\text{pool}}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\tau^{\text{pool}}).\end{aligned} \tag{38}$$

The federated variance $\hat{\mathbf{V}}^{\text{fed}}$ from sample size weighting in Section 3.3.2 satisfies

$$\hat{\mathbf{V}}_\tau^{\text{fed}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \hat{\mathbf{V}}_\tau^{(k)} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \hat{\mathbf{V}}_\tau^{(k)} \xrightarrow{p} \sum_{k=1}^{D} p_k \mathbf{V}_\tau^{(k)} = \mathbf{V}_\tau^{\text{pool}},$$

where we use the property that $\hat{\mathbf{V}}_\tau^{(k)} \xrightarrow{p} \mathbf{V}_\tau^{(k)}$ from Lemma 2.

For the case where $\phi(\mathbf{X}_i, W_i, Y_i)$ is the same across data sets, we have $\mathbf{V}_\tau^{\text{pool}} \equiv \mathbf{V}_\tau^{(k)} = \mathbf{V}_\tau$ for

all k and for some $\mathbf{V}_\tau$. Then, the federated variance $\hat{\mathbf{V}}_\tau^{\text{fed}}$ from sample size weighting in Section 3.3.2 satisfies

$$\hat{\mathbf{V}}_\tau^{\text{fed}} = \Big(\sum_{k=1}^{D} \big(\hat{\mathbf{V}}_\tau^{(k)}\big)^{-1} \Big)^{-1} \xrightarrow{p} \mathbf{V}_\tau.$$

The federated treatment effect $\hat{\tau}_{\text{aipw}}^{\text{fed}}$ from inverse variance weighting in Section 3.3.1 satisfies

$$\begin{aligned}
n_{\text{pool}}^{1/2}\big(\hat{\tau}_{\text{aipw}}^{\text{fed}} - \tau_0\big) =& n_{\text{pool}}^{1/2} \Big(\sum_{k=1}^{D} (\hat{\mathbf{V}}_\tau^{(k)})^{-1} \Big)^{-1} \Big(\sum_{k=1}^{D} (\hat{\mathbf{V}}_\tau^{(k)})^{-1} (\hat{\tau}_{\text{aipw}}^{(k)} - \tau_0) \Big) \\
=& n_{\text{pool}}^{1/2} \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} (\hat{\tau}_{\text{aipw}}^{(k)} - \tau_0) + o_p(1) \\
=& n_{\text{pool}}^{1/2} \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \cdot \frac{1}{n_k} \sum_{i=1}^{n_k} \phi^{\text{pool}}(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\tau^{\text{pool}}),
\end{aligned}$$

where the second equality uses Eq. (37).

For both cases, $\hat{\tau}_{\text{aipw}}^{\text{fed}}$ is asymptotically normal, and $\hat{\mathbf{V}}_\tau^{\text{fed}}$ is consistent. Then, from Slutsky's theorem, we have

$$\begin{aligned}
n_{\text{pool}}^{1/2} (\hat{\mathbf{V}}_\tau^{\text{fed}})^{-1/2} (\hat{\tau}_{\text{aipw}}^{\text{pool}} - \tau_0) \xrightarrow{d} & \mathcal{N}(0,1) \\
n_{\text{pool}}^{1/2} (\hat{\mathbf{V}}_\tau^{\text{pool}})^{-1/2} (\hat{\tau}_{\text{aipw}}^{\text{fed}} - \tau_0) \xrightarrow{d} & \mathcal{N}(0,1) \\
n_{\text{pool}}^{1/2} (\hat{\mathbf{V}}_\tau^{\text{fed}})^{-1/2} (\hat{\tau}_{\text{aipw}}^{\text{fed}} - \tau_0) \xrightarrow{d} & \mathcal{N}(0,1).
\end{aligned}$$

□