Federated causal inference in heterogeneous observational data

Ruoxuan Xiong | Allison Koenecke | Michael Powell | Zhu Shen | Joshua T. Vogelstein | Susan Athey

1Department of Quantitative Theory and Methods, Emory University, Atlanta, Georgia, USA
2Department of Information Science, Cornell University, Ithaca, New York, USA
3Department of Mathematical Sciences, United States Military Academy, West Point, New York, USA
4Department of Biostatistics, Harvard University, Cambridge, Massachusetts, USA
5Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland, USA
6Graduate School of Business, Stanford University, Stanford, California, USA

Correspondence
Ruoxuan Xiong, Department of Quantitative Theory and Methods, Emory University, Atlanta, GA, USA.
Email: rxiongpk@gmail.com

Funding information
Defense Advanced Research Projects Agency, Grant/Award Number: FA8650-18-2-7834; Microsoft Research; Office of Naval Research, Grant/Award Number: N00014-19-1-2468

We are interested in estimating the effect of a treatment applied to individuals at multiple sites, where data is stored locally for each site. Due to privacy constraints, individual-level data cannot be shared across sites; the sites may also have heterogeneous populations and treatment assignment mechanisms. Motivated by these considerations, we develop federated methods to draw inferences on the average treatment effects of combined data across sites. Our methods first compute summary statistics locally using propensity scores and then aggregate these statistics across sites to obtain point and variance estimators of average treatment effects. We show that these estimators are consistent and asymptotically normal. To achieve these asymptotic properties, we find that the aggregation schemes need to account for the heterogeneity in treatment assignments and in outcomes across sites. We demonstrate the validity of our federated methods through a comparative study of two large medical claims databases.

KEYWORDS
causal inference, federated learning, multiple data sets, propensity scores

1 INTRODUCTION

In many settings, the same treatment is applied to populations in different environments, but data is stored separately for each environment. When the sample size in any one data set is too small to obtain precise estimates of treatment effects, it would often be beneficial, if possible, to use data across environments. However, the combination of individual-level data may be restricted by 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 pooling individual-level data. Methods that accomplish this while sharing only aggregate data are referred to as “federated” learning methods. In this article, we develop federated learning methods tailored to the problem of causal inference. The methods allow
for heterogeneous treatment effects and heterogeneous outcome models across data sets, and adjust for the imbalance in covariate distributions between treated and control samples. These methods provide treatment effect estimation and inference, that are shown to perform as well asymptotically as if the data sets were combined.

A motivating example for these methods is from Koenecke et al., who study two separate medical claims data sets, MarketScan and Optum. The two data sets are noticeably different: the data from Optum has more elderly patients and covers more years than the data from MarketScan. They found evidence from both 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. However, existing federated methods are insufficient to draw inferences on the drug effect, while accounting for the heterogeneity in populations between treated and control groups and across two separate data sets.

In this article, we propose two main categories of federated inference methods to address this problem. One category is based on the inverse propensity-weighted maximum likelihood estimator (IPW-MLE),† The other one is based on the augmented inverse propensity weighted (AIPW) estimator. Our federated methods only use summary statistics of each data set and aim to estimate the parameters, such as average treatment effects, on the combined, individual-level data. Our methods provide point estimates and confidence intervals of these parameters that are asymptotically the same as if individual-level data were combined. We focus on IPW-MLE and AIPW for two main reasons. First, both estimators use propensity scores to balance covariate distributions between treated and control groups. Second, both estimators enjoy the double robustness property,‡ that are robust to the misspecification of one of the propensity and outcome models. As a building block, we propose a supplementary category of federated methods based on MLE for the estimation of either propensity or outcome model, and used as the inputs for the two main categories.

We make four contributions in developing federated inference methods. First, we identify the conditions that need to be considered in federation for valid inference, such as the stability of propensity and outcome models across data sets. Our federated inference methods are then designed to vary with these conditions. Second, to support the validity of inference, we develop inferential theory for all of our federated methods. Our federated methods achieve the optimal convergence rate in the estimation of average treatment effects and other parameters of interest. Third, our federated methods are communication-efficient. We show one-way and one-time sharing of carefully constructed summary statistics is sufficient to obtain consistent federated estimators. Fourth, for IPW-MLE, the estimation error in the propensity model carries over to the estimation of the outcome model,‡‡ which is often overlooked in practice, such as the standard SvYglm package in R.† Our federated IPW-MLE explicitly accounts for this estimation error.

Our federated methods are particularly relevant when separate data sets have heterogeneous populations with heterogeneous treatment assignment and outcome models. This is the setting where conventional pooling methods, such as inverse variance weighting (IVW), can fail.§ Let us revisit the example in Koenecke et al. We first estimate the effect of alpha blockers by IPW logistic regression on each data set. We then combine the estimated effects by IVW and by our federated IPW-MLE across data sets. As shown in Figure 1, the federated coefficient of alpha blockers from IVW lies outside of the interval defined by coefficients estimated on two separate data sets. This observation is counterintuitive as we expect the federated coefficient to measure the average effect of alpha blockers for patients in two data sets. In contrast, the federated coefficient from our proposed method lies between the coefficients estimated separately on two data sets, which makes more sense than IVW.

Our work is related to multiple streams of literature which aim to learn and analyze data from multiple sources, including streams from biostatistics, data mining, and federated learning. Most studies in data mining and federated learning focus on estimating a centralized model, mostly through an iterative approach while preserving privacy, without considering inference. In contrast, our federated methods are non-iterative and are supported by asymptotic theory.** Studies that provide inference are mostly concentrated in biostatistics. Specifically, early studies in meta-analysis and meta-regression analysis provide inference, but largely center around combining randomized controlled trials, and a typically used pooling approach is IVW. Recently, a growing number of studies develop privacy-preserving methods to provide inference by pooling aggregate data across multiple studies: most of them are tailored to specific parametric models, including linear models, logit models, Poisson models, Cox models, and generalized linear models, while Jordan et al. and Duan et al. consider the efficient pooling of the more general MLE. Among these studies, only Toh et al. and Shu et al. account for nonrandom treatment assignments by using propensity scores, though the asymptotic theory is lacking. In contrast, we provide federated methods for a general class of parametric models that adjust for nonrandom treatment assignments and are supported by asymptotic theory.

Our work is most closely related to the recent studies of privacy-preserving methods for causal inference by Vo et al. and Han et al. Vo et al. estimate treatment effects by modeling potential outcomes by Gaussian processes. Han et al. propose to estimate treatment effects for target populations by adaptively and optimally weighing source
FIGURE 1 Coefficient of the exposure to alpha blockers. This figure shows the estimated coefficient and its 95% confidence interval of the exposure to alpha blockers in a logit outcome model, where the outcome indicates whether the patient with acute respiratory distress (ARD) received mechanical ventilation and then had in-hospital death. We use IVW and our federated IPW-MLE to estimate the coefficient of alpha blockers on the combined data of MarketScan and Optum. The estimated coefficient from our federated IPW-MLE is more credible than that from IVW, because our federated coefficient lies between the interval defined by estimated coefficients on MarketScan and Optum, while the coefficient from IVW does not. See Section 5 for more details.

TABLE 1 A summary of matrices in the asymptotic variance of MLE and IPW-MLE.

| Matrix | Expression | Matrix | Expression |
|--------|------------|--------|------------|
| $\mathbf{A}_\beta$ | $\mathbb{E} \left[ \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right]$ | $\mathbf{A}_\gamma$ | $\mathbb{E} \left[ \frac{\partial \log f(y|x;\gamma)}{\partial \gamma} \right]$ |
| $\mathbf{B}_\beta$ | $\mathbb{E} \left[ \frac{\partial \log f(y|x;\beta)}{\partial \beta} \left( \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right)^\top \right]$ | $\mathbf{B}_\gamma$ | $\mathbb{E} \left[ \frac{\partial \log f(y|x;\gamma)}{\partial \gamma} \left( \frac{\partial \log f(y|x;\gamma)}{\partial \gamma} \right)^\top \right]$ |
| ATE weighting $m_{ATE} = \frac{w}{e_\gamma(x)} + \frac{1-w}{1-e_\gamma(x)}$ | ATT weighting $m_{ATT} = w_1 + \frac{e_\gamma(x)}{1-e_\gamma(x)}(1-w_1)$ |
| $\mathbf{A}_{\beta,\omega}$ | $\mathbb{E} \left[ w + \frac{1-w}{1-e_\gamma} \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right]$ | $\mathbf{A}_{\beta,\omega}$ | $\mathbb{E} \left[ w + \frac{e_\gamma(x)}{1-e_\gamma}(1-w) \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right]$ |
| $\mathbf{D}_{\beta,\omega}$ | $\mathbb{E} \left[ \frac{w}{e_\gamma} \frac{1-w}{1-e_\gamma} \frac{\partial \log f(y|x;\beta)}{\partial \beta} \left( \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right)^\top \right]$ | $\mathbf{D}_{\beta,\omega}$ | $\mathbb{E} \left[ \frac{w}{e_\gamma} \frac{e_\gamma(x)}{1-e_\gamma}(1-w) \frac{\partial \log f(y|x;\beta)}{\partial \beta} \left( \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right)^\top \right]$ |
| $\mathbf{C}_{\beta,\omega}$ | $\mathbb{E} \left[ \frac{w}{e_\gamma} \frac{1-w}{1-e_\gamma} \frac{\partial \log f(y|x;\beta)}{\partial \beta} \left( \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right)^\top \right]$ | $\mathbf{C}_{\beta,\omega,1}$ | $\mathbb{E} \left[ \frac{1-w}{1-e_\gamma} \frac{\partial \log f(y|x;\beta)}{\partial \beta} \left( \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right)^\top \right]$ |
| $\mathbf{C}_{\beta,\omega,2}$ | $\mathbb{E} \left[ \frac{w}{e_\gamma} \frac{1-w}{1-e_\gamma} \frac{\partial \log f(y|x;\beta)}{\partial \beta} \left( \frac{\partial \log f(y|x;\beta)}{\partial \beta} \right)^\top \right]$ |

Note: In the definitions of these matrices, $e_\gamma(x)$ denotes $e_\gamma(x) = \alpha(x, \gamma)$ by a slight abuse of notation.

In contrast, our federated inference methods focus on treatment effects and other parameters of interest defined on the combined data, as opposed to on specific target data as in Han et al.\textsuperscript{28,29}

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 define the target parameters in our federated estimators in Section 2.2. We then review three widely used estimators (MLE, IPW-MLE, AIPW) on which our federated estimators are built in Section 2.3. Next, we list the covariate and model conditions that need to be considered in the federation in Section 2.4. Finally, in Section 2.5, we state the three weighting methods to aggregate information in our federated estimators. All the matrices in the asymptotic variance of MLE and IPW-MLE are summarized in Table 1.

2.1 | Model setup

Suppose we have $D$ data sets, where $D$ is finite. Suppose data set $k \in \{1, \ldots, D\}$ has $n_k$ observations $(X_i^{(k)}, Y_i^{(k)}, W_i^{(k)}) \in \mathcal{X} \times \mathbb{R} \times \{0, 1\}$ that are drawn i.i.d. from some distribution $\mathbb{P}^{(k)}$. Here, $i \in \{1, \ldots, n_k\}$ indexes the subjects (e.g., patients), populations, accounting for the risk of negative transfer when source and target populations are heterogeneous.
$\mathbf{X}_i^{(k)}$ is a vector of $d_k$ observed covariates, $Y_i^{(k)}$ is the outcome of interest, $W_i^{(k)}$ is the treatment assignment, and $\mathcal{X}_k \subseteq \mathbb{R}^{d_k}$. Both the types and the number of covariates can vary with data sets. Let $n_{\text{pool}} = \sum_{k=1}^{K} n_k$ be the total number of observations. Here we study the setting where each data set has many observations, that is, $n_k$ is large for all $k$. We assume the population fraction of observations in data set $k$, that is, $p_k = \lim n_k / n_{\text{pool}}$, exists, and is bounded away from 0 and 1.

Under the Neyman-Rubin potential outcome model and the stable unit treatment value assumption, let $Y_i^{(k)}(1)$ be the outcome of subject $i$ if it is assigned treatment, and let $Y_i^{(k)}(0)$ be the outcome for the opposite case. For each data set $k$, suppose the following standard unconfoundedness assumption \(^{37}\) holds

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

and the following overlap assumption \(^{37}\) for the propensity score $e^{(k)}(x) = pr(W_i^{(k)} = 1 | \mathbf{X}_i^{(k)} = x)$ holds

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

for some $\eta > 0$. For each data set $k$, we define the average treatment effect (ATE), denoted as $r_{\text{ate}}^{(k)}$, and average treatment effect on the treated (ATT), denoted as $r_{\text{att}}^{(k)}$, as follows

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

### 2.1.1 Parametric models

In this article, we focus on parametric outcome and propensity models stated in Conditions 1 and 2 below. This is motivated by the common use of parametric outcome models in medical applications, for example, the use of logistic regression for estimating the odds ratio in epidemiological studies, Cox regression for survival analysis in clinical trials, and generalized linear models (GLM) for assessing medical costs. In addition, parametric models, such as logit models, are also commonly used to estimate propensity scores. \(^{36}\) The estimated parametric outcome and/or propensity model can also be used as the input in the estimation of the ATE and ATT.

**Condition 1** (Parametric outcome model). For any data set $k$, the conditional density function of outcome $y$ on $x$ and $w$ follows a parametric model, denoted as $f^{(k)}_0(y|x, w, \beta)$ with the true parameter values to be $\beta^{(k)}_0$.

**Condition 2** (Parametric propensity model). For any data set $k$, the conditional treatment probability $pr(w = 1|x)$ follows a parametric model, denoted as $e^{(k)}_0(x, \gamma)$, with the true parameter values to be $\gamma^{(k)}_0$.

Given Conditions 1 and 2, we can estimate the outcome and propensity models by maximizing the (weighted) likelihood function. Since the parametric models $f^{(k)}_0(y|x, w, \beta)$ and $e^{(k)}_0(x, \gamma)$ are unknown a priori, the family of distributions chosen in the estimation of outcome and propensity models, denoted as $f^{(k)}(y|x, w, \beta)$ and $e^{(k)}(x, \gamma)$, may or may not contain the true structure, $f^{(k)}_0(y|x, w, \beta)$ and $e^{(k)}_0(x, \gamma)$. Our federated estimators account for the possibility of model misspecification. We further discuss when the particular parameters of interest, for example, ATE or ATT, on the combined data can still be consistently estimated by federated estimators in the presence of misspecification.

### 2.2 Target parameters

In this subsection, we define the target parameters that our federated methods aim to estimate. Throughout this article, the superscript “$(k)$” in a notation denotes an object estimated using data set $k$; the superscript “cb” denotes an object on the combined, individual-level data; and the superscript “fed” denotes a federated estimator.

The target parameters are defined on the combined data that concatenate individual data across $D$ data sets together. The first set of target parameters are the parameters in the true conditional outcome density $f^{(cb)}_0(\cdot)$ on the combined data, denoted as $\beta^{(cb)}_0$, where $f^{(cb)}_0(\cdot)$ is defined as

$$f^{(cb)}_0(Y_i^{(k)}|X_i^{(k)}, W_i^{(k)}, \beta^{(cb)}_0) := \prod_{j=1}^{K} \left[ f^{(j)}_0(Y_i^{(j)}|X_i^{(j)}, W_i^{(j)}, \beta^{(j)}_0) \right]^{I(j=k)}, \quad \forall k,$$
that equals the true conditional outcome density of data set \( k \) when the observation is from data set \( k \). \( \beta^{cb}_0 \) is defined as the union of \( \beta^{(1)}_0, \ldots, \beta^{(K)}_0 \). For example, if \( \beta^{(1)}_0 = \cdots = \beta^{(K)}_0 \), then \( \beta^{cb}_0 = \beta^{(k)}_0 \) for any \( k \); if \( \beta^{(1)}_0, \ldots, \beta^{(K)}_0 \) is completely different from one another, then \( \beta^{cb}_0 = (\beta^{(1)}_0, \ldots, \beta^{(K)}_0) \).

The second set of target parameters are the parameters in the true propensity \( e^{cb}_0 (\cdot) \) on the combined data, denoted as \( r^{cb}_0 \), where \( e^{cb}_0 (\cdot) \) is defined as

\[
e^{cb}_0 (W^{(k)}_i | X^{(k)}_i, r^{cb}_0) := \prod_{j=1}^{K} \left[ e^{(j)}_0 (W^{(j)}_i | X^{(j)}_i, r^{(j)}_0) \right]^{\mathbb{I}(j = k)}, \quad \forall k,
\]

that equals the true propensity of data set \( k \) when the observation is from data set \( k \). Similar to \( \beta^{cb}_0 \), \( r^{cb}_0 \) is defined as the union of \( r^{(1)}_0, \ldots, r^{(K)}_0 \).

The third set of target parameters are the ATE and ATT on the combined data, denoted as \( \tau^{cb}_{\text{ate}} \) and \( \tau^{cb}_{\text{att}} \), and are defined as

\[
\tau^{cb}_{\text{ate}} := \sum_{k=1}^{D} p_k \tau^{(k)}_{\text{ate}}, \quad \tau^{cb}_{\text{att}} := \sum_{k=1}^{D} p_k \tau^{(k)}_{\text{att}},
\]

where \( \tau^{cb}_{\text{ate}} \) and \( \tau^{cb}_{\text{att}} \) are the averages of \( \tau^{(k)}_{\text{ate}} \) and \( \tau^{(k)}_{\text{att}} \) weighted by \( p_k \), and \( p_k \) is the population fraction of observations in data set \( k \). Both \( \tau^{cb}_{\text{ate}} \) and \( \tau^{cb}_{\text{att}} \) do not depend on the sample size.

If data sets can be combined at the individual level, then the standard approaches for a single data set (as reviewed in Section 2.3 below) are applicable to estimate and draw inference on these target parameters. However, when data sets cannot be combined at the individual level, standard approaches are not applicable.

We develop federated inference methods for these target parameters that only use aggregate information from each data set. The federated inference methods consist of both point and variance estimators of target parameters, thus allowing for the construction of confidence intervals of target parameters. These confidence intervals can be narrower than those obtained from a single data set. When treatment assignments are randomized, our federated methods include classical approaches such as IVW in meta-analysis, whereas when they are nonrandom, our federated estimators adjust for selection bias.

Note that in some settings, such as those in transfer learning, the target parameters of interest are defined on a specific target data set. Other data sets are used to improve the estimation efficiency of target data. In these settings, if propensity and outcome models are stable (defined in Conditions 4 and 5 below), then our federated estimators continue to be valid; otherwise, we need to account for the discrepancy between supplementary and target data sets to avoid the negative transfer. See Han et al28 for more discussion.

### 2.3 Estimation methods for combined individual-level data

This subsection reviews MLE, IPW-MLE, and AIPW that could be used to estimate the target parameters in Section 2.2 when individual-level data could have been combined. As the individual data cannot be combined in practice, the estimators in this subsection are not feasible. In Section 3, we introduce our federated estimators that are designed to approximate the estimators in this section using only the summary statistics of each data set.

#### 2.3.1 MLE for model parameters

Under the parametric outcome model, we define the log-likelihood function of outcome conditional on covariates and treatment assignment on the combined data as

\[
\mathcal{L}_{\text{pool}} (\beta) = \sum_{k=1}^{D} \sum_{i=1}^{n_k} \log f(Y_i^{(k)} | X_i^{(k)} \cdot W_i^{(k)}, \beta),
\]

(2)
where $\mathcal{L}_{n_k}(\beta)$ is the log-likelihood function on data set $k$. Let $\hat{\beta}_{\text{mle}}^{cb}$ be the solution that maximizes the log-likelihood function $\mathcal{L}_{n_k}(\beta)$ and $\hat{\beta}_{\text{mle}}^{cb}$ is an estimator of $\beta^{cb}$. We can analogously use MLE to estimate the parameters in the parametric propensity model on the combined data.

2.3.2 IPW-MLE for model parameters and average treatment effects

An alternative approach to estimating parameters in the outcome model is to use IPW-MLE, which adjusts the log-likelihood function by inverse propensity scores to estimate the population mean when data is nonrandomly missing.

$$
\mathcal{L}_{n_k}(\beta, \hat{e}) = \sum_{k=1}^{D} \sum_{i=1}^{n_k} \mathcal{L}_{i,k}^{(\beta)} \log f(Y_i^{(k)}|X_i^{(k)}, W_i^{(k)}, \beta),
$$

where the subscript “$\hat{e}$” is the abbreviation of the estimated propensity on the combined data, $\mathcal{L}_{n_k}(\beta, \hat{e})$ is the weighted log-likelihood function on data set $k$, and $\mathcal{L}_{i,k}^{(\beta)}$ is the weight for unit $i$ that can be

$$
\mathcal{L}_{i,k}^{(\beta)} = \begin{cases} 
W_i^{(k)} / \hat{e}(X_i^{(k)}) + (1 - W_i^{(k)}) / (1 - \hat{e}(X_i^{(k)})) & \text{ATE weighting} \\
W_i^{(k)} + \hat{e}(X_i^{(k)}) (1 - W_i^{(k)}) / (1 - \hat{e}(X_i^{(k)})) & \text{ATT weighting}.
\end{cases}
$$

Let $\hat{\beta}_{\text{IPW-mle}}^{cb}$ be the estimator that maximizes the weighted log-likelihood $\mathcal{L}_{n_k}(\beta, \hat{e})$. This estimator can be used to estimate treated and control outcomes, and form a doubly robust estimator for ATE and ATT. See Appendix A.2 for more details.

2.3.3 AIPW for average treatment effects

We can estimate ATE on the combined data using the AIPW estimator

$$
\hat{\tau}_{\text{AIPW}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \cdot \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{\hat{\phi}(X_i^{(k)}, W_i^{(k)}, Y_i^{(k)})}{\hat{e}(X_i^{(k)})},
$$

that can be written as a weighted average of ATE across data sets by sample size, where $\hat{\phi}(\cdot)$ is the estimated score on the combined data and is defined as

$$
\hat{\phi}(x, w, y) = \hat{\mu}_1(x) - \hat{\mu}_0(x) + \frac{w}{\hat{e}(x)} (y - \hat{\mu}_1(x)) - \frac{(1 - w)}{1 - \hat{e}(x)} (y - \hat{\mu}_0(x)),
$$

and where $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ are estimated conditional treated and control outcome models on the combined data. If the estimand is ATT, then we can also use (4), but the estimated score $\hat{\phi}(\cdot)$ is defined as

$$
\hat{\phi}(x, w, y) = w (y - \hat{\mu}_1(x)) - \frac{\hat{e}(x)(1 - w)}{1 - \hat{e}(x)} (y - \hat{\mu}_0(x))
$$

AIPW has two prominent properties: doubly robustness and semiparametric efficiency.

2.4 Covariate and model considerations in federated estimators

In this subsection, we introduce the conditions that need to be considered in the federation to obtain valid point and variance estimators of target parameters.
Condition 3 (Known propensity score). For all data sets, the true propensity scores are known and used. When true propensity scores are known and used, then we do not need to federate propensity models in federated IPW-MLE.

Condition 4 (Stable propensity model). The set of covariates and the parameters in the propensity model are the same for all data sets, that is, $\gamma_0^{(j)} = \gamma_0^{(k)}$ for any $j$ and $k$.

Condition 5 (Stable outcome model). The set of covariates and the parameters in the outcome model are the same for all data sets, that is, $\beta_0^{(j)} = \beta_0^{(k)}$ for any $j$ and $k$.

Condition 6 (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 $\mathbb{P}^{(j)}(x) = \mathbb{P}^{(k)}(x)$ for any two data sets $j$ and $k$.

We refer to data sets as being “heterogeneous” in settings where either Condition 4, 5, or 6 is violated. If Condition 5 holds (similarly for Condition 4), then the parameters on the combined data $\beta_{cb}^{0}$ equals $\beta_{0}^{(k)}$ for any $k$; otherwise, we partition the parameters $\beta^{(k)} = (\beta_{s}, \beta_{uns}^{(k)})$ into shared parameters $\beta_{s}$ and dataset-specific parameters $\beta_{uns}^{(k)}$ for any $k$, and define the parameters on the combined data as $\beta_{cb}^{0} = (\beta_{s}, \beta_{uns}^{(1)}, \beta_{uns}^{(2)}, \ldots, \beta_{uns}^{(D)})$.\textsuperscript{38} For example, $\beta_{s}$ could include the parameters of interest, such as the treatment coefficient that we want to precisely estimate; $\beta_{uns}^{(k)}$ could include nuisance parameters, such as the age coefficient in our empirical study.\textsuperscript{34} Note that choosing the partition generally encompasses a tradeoff between efficiency and robustness to model misspecification. See Section 3.1.2 for more discussion, and Section 3.4 for practical guidance on choosing the partition.

2.5 Three weighting methods

We list the three weighting methods used in our federated estimators. The choice of weighting methods in each federated estimator is based on the functional form of the corresponding estimator for a single data set, as shown in Section 3, and ensures that the federated estimators can be consistent, as shown in Section 4.

2.5.1 Hessian weighting

Hessian weighting is used to estimate target parameters $\beta_{cb}^{0}$ and $\gamma_{0}^{cb}$ in the outcome and propensity models, and is defined as

$$\hat{\beta}_{fed}^{0} = \left( \sum_{k=1}^{D} \hat{H}_{\beta}^{(k)} \right)^{-1} \left( \sum_{k=1}^{D} \hat{H}_{\beta}^{(k)} \hat{\beta}^{(k)} \right), \text{ where } \hat{H}_{\beta}^{(k)} = \frac{\partial^2 l_{nk}^{(k)}}{\partial \beta^{(k)} \partial \beta^{(k)}}.$$  \hfill (7)

for parameters in the outcome model. For the propensity model, we just replace $\hat{\beta}^{(k)}$ by $\hat{\gamma}^{(k)}$ and $\hat{H}_{\beta}^{(k)}$ by $\hat{H}_{\gamma}^{(k)}$ in (7).

2.5.2 Sample size weighting

Sample size weighting is used to obtain variance estimators (see more details in Tables 2–4), and is used to estimate ATE and ATT under unstable propensity or outcome models. For some generic scalar or matrix $M$, we refer to sample size weighting as

$$M_{fed}^{0} = \sum_{k=1}^{D} \frac{n_k}{n_{pool}} M^{(k)}, \text{ where } n_{pool} = \sum_{k=1}^{D} n_k.$$  \hfill (8)
| Description | Assume stable outcome model (MLE #1) | Assume unstable outcome model (MLE #2) |
|-------------|-------------------------------------|--------------------------------------|
| Stable outcome model | Yes | No |
| Parameter $\beta$ federation | $(\sum_{k=1}^{D} \hat{H}_{\beta}^{(k)})^{-1} (\sum_{k=1}^{D} \hat{H}_{\beta}^{(k)} \hat{\beta}^{(k)})$ | $(\sum_{k=1}^{D} \hat{H}_{\beta}^{(k)})^{-1} (\sum_{k=1}^{D} \hat{H}_{\beta}^{(k)} \hat{\beta}^{(k)})$ |
| Variance $\mathbf{V}_{\beta}$ federation | Sample size weighting $\hat{A}_{\beta}^{(k)}$ and $\hat{B}_{\beta}^{(k)}$ in $\mathbf{V}_{\beta} = \hat{A}_{\beta}^{-1} \hat{B}_{\beta} \hat{A}_{\beta}^{-1}$ | Sample size weighting $\hat{A}_{\beta}^{(k)} \hat{B}_{\beta}^{(k)}$ and $\hat{B}_{\beta}^{(k)}$ in $\mathbf{V}_{\beta} = \hat{A}_{\beta}^{-1} \hat{B}_{\beta} \hat{A}_{\beta}^{-1}$ |
| Asymptotic results | Theorem 1 | |

Note: This table also holds for the propensity model. The second row corresponds to Condition 5. $\hat{H}_{\beta}^{(k)}$ denotes the estimated Hessian. $\hat{A}_{\beta}$ and $\hat{B}_{\beta}$ are defined in Table 1. $\hat{H}_{\beta}^{(k)}$ increases with sample size $n_k$, while $\hat{A}_{\beta}$ and $\hat{B}_{\beta}$ do not. For a generic vector or matrix $\mathbf{x}$, $\mathbf{x}^{\text{pad}}$ denotes $\mathbf{x}$ padded with zeros.

| Description | Assume known propensity and stable outcome model (IPW-MLE #1) | Assume stable propensity and stable outcome model (IPW-MLE #2) | Assume unstable propensity or unstable outcome model (IPW-MLE #3) |
|-------------|-------------------------------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|
| Stable propensity model | Yes or no | Yes | Yes or no |
| Stable outcome model | Yes | Yes | Yes or no |
| Parameter $\beta$ federation | (1) Estimate $\hat{\beta}^{(k)}$ using $\hat{\gamma}_0$; (2) Federate $\hat{\beta}^{(k)}$ by Hessian weighting | (1) Federate $\hat{\gamma}^{(k)}$ by Hessian weighting; (2) Estimate $\hat{\beta}^{(k)}$ using $\hat{\gamma}^{\text{fed}}$; (3) Federate $\hat{\beta}^{(k)}$ by Hessian weighting | Same federation procedure, but with $\hat{\gamma}^{\text{fed}(k)}$ and $\hat{H}_{\beta}^{\text{fed}(k)}$ if propensity models are unstable and estimated, and with $\hat{\beta}^{\text{pad}(k)}$ and $\hat{H}_{\beta}^{\text{pad}(k)}$ if outcomes models are unstable |
| Variance $\mathbf{V}_{\beta}$ federation | $\mathbf{V}_{\beta} = A_{\beta, m}^{-1} (D_{\beta, m} - M_{\beta, m, 0}) A_{\beta, m}^{-1}$, (1) Estimate $A_{\beta, m}^{-1}$ using $\hat{\beta}^{\text{fed}}$; (2) Federate $\hat{A}_{\beta, m}$ and $\hat{D}_{\beta, m}$ by sample size weighting | $\mathbf{V}_{\beta} = A_{\beta, m}^{-1} (D_{\beta, m} - M_{\beta, m, 0}) A_{\beta, m}^{-1}$, $M_{\beta, m, 0} = C_{\beta, m, 0} V_{\beta} C_{\beta, m, 0}^T$ for ATE weighting; $M_{\beta, m, 0} = C_{\beta, m, 0} V_{\beta} C_{\beta, m, 0}^T + C_{\beta, m, 2} V_{\beta} C_{\beta, m, 2}^T - C_{\beta, m, 2} V_{\beta} C_{\beta, m, 2}^T$ for ATT weighting; $\mathbf{V}_{\beta} = A_{\beta, m}^{-1} B_{\beta} A_{\beta, m}^{-1}$, (1) Estimate $A_{\beta, m}^{-1}$, $C_{\beta, m, 0}$, $D_{\beta, m}$, $A_{\beta}^{(k)}$, and $B_{\beta}^{(k)}$ using $\hat{\gamma}^{\text{fed}}$ and $\hat{\beta}^{\text{fed}}$; (2) Federate $\hat{A}_{\beta, m}$, $\hat{C}_{\beta, m}$, $\hat{D}_{\beta, m}$, $\hat{A}_{\beta}^{(k)}$, and $\hat{B}_{\beta}^{(k)}$ by sample size weighting | Same federation procedure, but with $\hat{\gamma}^{\text{fed}(k)}$, $\hat{A}_{\beta, m}^{\text{pad}(k)}$, $\hat{C}_{\beta, m}^{\text{pad}(k)}$, and $\hat{B}_{\beta}^{\text{pad}(k)}$ if propensity models are unstable and estimated, and with $\hat{\beta}^{\text{pad}(k)}$, $\hat{A}_{\beta, m}^{\text{pad}(k)}$, $\hat{D}_{\beta, m}^{\text{pad}(k)}$, and $\hat{C}_{\beta, m}^{\text{pad}(k)}$ if outcomes models are unstable |
| Asymptotic results | Theorem 2 | | |

Note: The second and third rows correspond to Conditions 4 and 5. “Yes or no” means that the solution does not vary with whether the condition is satisfied or not. The definitions of $A_{\beta, m}, D_{\beta, m}, C_{\beta, m}, C_{\beta, m, 0}, C_{\beta, m, 2}, A_{\beta}$, and $B_{\beta}$ can be found in Table 1. When the propensity model is estimated (Condition 3 is violated), the coefficient federation procedure is the same for all scenarios, but is simplified when the true propensity is used (Condition 3 holds). The variance federation procedure varies with whether the true propensity is used and whether ATE or ATT weighting is used. The definitions of ATE and ATT weighting can be found in Section 2.3.2. For a generic vector or matrix $\mathbf{x}$, $\mathbf{x}^{\text{pad}}$ denotes $\mathbf{x}$ padded with zeros.
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 propensity model                  | Yes                                                         | Yes or no                                                     |
| Stable outcome model                     | Yes                                                         | Yes or no                                                     |
| Stable covariate distribution            | Yes                                                         | Yes or no                                                     |
| ATE or ATT federation                    | (1) Federate \( \hat{\beta}^{(k)} \) (and \( \hat{\gamma}^{(k)} \) if necessary) by Hessian weighting; (2) Estimate \( \tau^{(k)} \) using \( \hat{\beta}^{\text{fed}} \) and \( \hat{\gamma}^{\text{fed}} \) (or \( \gamma^{(k)} \) if known); (3) Federate \( \hat{\tau}^{(k)} \) by inverse variance weighting | (1) Estimate \( \tau^{(k)} \) using \( \hat{\beta}^{(k)} \) and \( \hat{\gamma}^{(k)} \) (or \( \gamma^{(k)} \) if known); (2) Federate \( \hat{\tau}^{(k)} \) by sample size weighting |
| Variance \( \nu_f \) federation         | Inverse variance weighting                                  | Sample size weighting                                          |

Results

The second to fourth rows correspond to Conditions 4–6.

FIGURE 2  Flowchart for federated MLE. See Section 3.4 for practical guidance on determining whether the outcome model is stable.

2.5.3  Inverse variance weighting

IVW is used to estimate ATE and ATT and their variance under stable propensity and outcome models. For some generic point estimator \( \hat{\nu} \), we refer to IVW as

\[
\hat{\nu}^{\text{fed}} = \left( \sum_{k=1}^{D} \left( \text{Var}(\hat{\nu}^{(k)}) \right)^{-1} \right)^{-1} \left( \sum_{k=1}^{D} \left( \text{Var}(\hat{\nu}^{(k)}) \right)^{-1} \nu^{(k)} \right),
\]

(9)

\[
\text{Var}(\hat{\nu}^{\text{fed}}) = n_{\text{pool}} \left( \sum_{k=1}^{D} \left( \text{Var}(\hat{\nu}^{(k)}) \right)^{-1} \right)^{-1},
\]

(10)

where \( \text{Var}(\hat{\nu}) \) is the variance of \( \hat{\nu} \), and \( \text{Var}(\hat{\nu}) \) is the variance of \( \hat{\nu} \) multiplied by the sample size.

3  FEDERATED ESTIMATORS

In this section, we introduce three categories of federated inference methods that consist of both point and variance estimators of target parameters in Section 2.2. These three categories are based on MLE, IPW-MLE, and AIPW, respectively. For each category, we start with the simple case in which the propensity and outcome models are stable. We refer to the federated estimators in this case as restricted federated estimators. Next, we consider the more challenging case in which at least one of the propensity and outcome models is unstable. The federated estimators for this case are referred to as unrestricted federated estimators, which are built on the corresponding restricted federated estimators.

Figures 2–4 show the flowcharts of our federated inference methods under different conditions. Tables 2–4 provide the details of our federated methods.
3.1 Federated MLE

We introduce our federated MLE using the outcome model, where the target parameter is $\beta_{cb}$. However, our federated MLE is also applicable to the propensity model.

3.1.1 Restricted federated MLE for stable models (Condition 4 or 5 holds)

When outcome models are stable (i.e., $\beta_{cb} = \beta^{(k)}$ for all $k$), we can use the restricted federated MLE for $\beta_{cb}$. Let $\hat{\beta}_{mle}^{fed}$ be the federated point estimator that is obtained by first applying MLE on each data set $k$ to estimate parameter $\beta^{(k)}$, and then using Hessian weighting in (7) to combine estimated parameters across all data sets.

We propose this federated estimator based on the objective of satisfying the first-order condition of MLE. When we use Hessian weighting, this objective can be satisfied with the key steps outlined below:

$$
\frac{\partial}{\partial \beta} \sum_{k=1}^{D} \mathcal{E}_{n_k}(\hat{\beta}_{mle}^{fed}) = \sum_{k=1}^{D} \frac{\partial \mathcal{E}_{n_k}(\beta_0)}{\partial \beta} + \left( \sum_{k=1}^{D} H_{\beta}^{(k)} \right) (\hat{\beta}_{mle}^{fed} - \beta_0)
$$

$$
= \sum_{k=1}^{D} \frac{\partial \mathcal{E}_{n_k}(\beta_0)}{\partial \beta} + \sum_{k=1}^{D} H_{\beta}^{(k)} (\hat{\beta}_{mle}^{fed} - \beta_0) \quad \text{(follow that } \hat{\beta}_{mle}^{fed} \text{ is obtained by Hessian weighting)}
$$

$$
= \sum_{k=1}^{D} \frac{\partial \mathcal{E}_{n_k}(\hat{\beta}_{mle}^{(k)})}{\partial \beta} = 0 \quad \text{(follow that gradient at } \hat{\beta}_{mle}^{(k)} \text{ is zero for all } k).
$$

Our federated variance estimator is obtained via a two-step procedure. First, we estimate the terms in the robust variance formula, $A_\beta$ and $B_\beta$ (see Table 1 for the definition), on each data set. Let $\hat{A}_{\beta}^{(k)}$ and $\hat{B}_{\beta}^{(k)}$ be the estimators on data.
set $k$. Second, we obtain the federated variance using sample size weighting**

$$
\mathbf{V}_\beta^{\text{fed}} = (\mathbf{A}_\beta^{\text{fed}})^{-1} \cdot \mathbf{B}_\beta^{\text{fed}} \cdot (\mathbf{A}_\beta^{\text{fed}})^{-1},
$$

where

$$
\mathbf{A}_\beta^{\text{fed}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \mathbf{A}_\beta^{(k)} \quad \text{and} \quad \mathbf{B}_\beta^{\text{fed}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \mathbf{B}_\beta^{(k)}. \tag{11}
$$

This federated variance uses the robust variance formula and is, therefore, robust to outcome model misspecification.43

We use sample size weighting here based on the property that $\mathbf{A}_\beta$ and $\mathbf{B}_\beta$ on the combined data equal the weighted average of the corresponding matrices on individual data sets by sample size.

### 3.1.2 Unrestricted federated MLE for unstable models (Condition 4 or 5 is violated)

Our unrestricted federated MLE is conceptually similar to our restricted federated MLE, but additionally handles the instability of parameters across datasets. Specifically, our unrestricted estimator only combines the shared parameters across data sets and leaves the dataset-specific parameters as they are in the federation. The key to treating shared and dataset-specific parameters differently is to use a zero-padding technique.

Specifically, for each data set $k$, we pad $\beta^{(k)}$ with zeros so that the padded $\beta^{\text{pad}(k)}$, denoted as $\beta^{\text{pad}(k)}$, is aligned with $\beta^{cb} = (\beta_s, \beta^{(1)}_{\text{uns}}, \beta^{(2)}_{\text{uns}}, \ldots, \beta^{(D)}_{\text{uns}})$. We similarly pad each matrix on data set $k$ so that it is aligned with the corresponding matrix on the combined data. Below we provide an example of zero-padding $\beta^{(1)}$ and $H^{(1)}_\beta$ for data set $k = 1$:

$$
\beta^{\text{pad}(1)} = \begin{pmatrix}
\beta_s \\
\beta^{(1)}_{\text{uns}} \\
0
\end{pmatrix}, \quad H^{\text{pad}(1)}_\beta = \begin{pmatrix}
H_{\beta,s,s}^{(1)} & H_{\beta,s,\text{uns}}^{(1)} & 0 \\
H_{\beta,\text{uns},s}^{(1)} & H_{\beta,\text{uns},\text{uns}}^{(1)} & 0 \\
0 & 0 & 0
\end{pmatrix}. \tag{12}
$$

The zero-padding of other vectors and matrices for other $k$ is conceptually the same. The unrestricted point and variance estimator essentially apply the restricted point and variance estimator to the padded parameters and matrices. In this way, the unrestricted estimator only federates the shared parameters.

Note that it is possible to treat some parameters as dataset-specific parameters even though they are stable. This approach does not affect the consistency of the federated estimator; however, as the number of parameters on the combined data increases, the federated estimator is weakly less efficient than that using the most parsimonious specification, as stated in the following proposition. See Table B5 in Appendix B.6 for a numerical example.

**Proposition 1.** Suppose $Y_i$ follows a generalized linear model that is stable across data sets (Condition 5 holds).

If we use unrestricted federated MLE with a flexible outcome model specification on combined data (i.e., $\beta^{cb}$ has a higher dimension than the most parsimonious specification), then we get a weakly less efficient estimate of $\beta_s$ than that from restricted federated MLE.

### 3.2 Federated IPW-MLE

The target parameter of our federated IPW-MLE is $\beta^{cb}$ in the outcome model on the combined data. As IPW-MLE uses propensity scores, we need to account for whether the propensity scores are known or estimated. If they are estimated, then our federated IPW-MLE also estimates and federates the propensity models.

#### 3.2.1 Restricted federated IPW-MLE for stable models (Conditions 4 and 5 hold)

Let $\hat{\beta}^{\text{fed}}_{\text{ipw-mle}}$ be our restricted federated point estimator for $\beta^{cb}$ obtained via a three-step procedure. First, if the propensity scores are unknown, we use restricted MLE to estimate the parameters in the propensity model on the combined data.
and obtain the federated propensity scores; otherwise, skip this step. Second, we use IPW-MLE with federated propensity scores to estimate $\beta^{(k)}$ on each data set $k$. Third, we combine estimated $\beta^{(k)}$ by Hessian weighting to obtain $\hat{\beta}^{\text{fed}}_{\text{ipw-mle}}$. Similar to federated MLE, this federated point estimator is designed to satisfy the first-order condition of IPW-MLE.

The federated variance estimator of IPW-MLE is designed based on the variance formula of IPW-MLE in Lemma 1 in Section 4.2 for a single data set. For every term in the variance formula, we estimate it on each data set. We combine the estimated terms across data sets by sample size weighting, and plug the sample size weighted terms into the variance formula to obtain the federated variance. The procedure is conceptually similar to that for MLE, but operates on a different variance formula. See Table 3 for more details.

3.2.2 Unrestricted federated IPW-MLE for unstable models (Condition 4 or 5 is violated)

Similar to unrestricted federated MLE, our unrestricted federated IPW-MLE only federates shared parameters in the propensity and outcome models, and leaves the dataset-specific parameters as they are in federation. We first pad the parameters and matrices on each data set with zeros to match the dimensionality of the corresponding parameters and matrices on the combined data. Then we apply restricted federated IPW-MLE to the zero-padded parameters and matrices to obtain point and variance estimates of the target parameter.

3.3 Federated AIPW estimator

Our federated AIPW estimates ATE or ATT on the combined data. The illustration of federated AIPW uses ATE as an example. The federation of ATT is conceptually the same.

3.3.1 Restricted AIPW estimator for stable models and stable covariate distributions (Conditions 4–6 hold)

As the AIPW estimator uses both outcome and propensity models, we need to federate both propensity and outcome models. When covariate distributions, propensity models, and outcome models are stable, we propose to use the restricted federated AIPW, which has three steps. First, we use federated MLE to obtain a federated propensity model and a federated outcome model. Second, we use AIPW with the federated propensity and outcome models to estimate ATE on each data set. Finally, we obtain the federated ATE by inverse variance weighting the estimated ATE on each data set, as in formula (9).

To obtain the federated variance, we first estimate the variance of the estimated ATE on each data set, and then use IVW to combine the estimated variances on all data sets together, as in formula (10).

Note that, under stable covariate distributions and stable propensity and outcome models, ATE and asymptotic variance of ATE are the same for all data sets. In this case, we can apply any weighting scheme to combine the estimated ATE together. We choose IVW because it has the smallest variance among all weighting schemes, as shown in Appendix A.5.

3.3.2 Unrestricted AIPW estimator for unstable models or unstable covariate distributions (either Condition 4, 5, or 6 is violated)

When either propensity model, outcome model, or covariate distribution is unstable, ATE may not be the same across data sets. For this case, we suggest using the unrestricted federated AIPW. For this unrestricted estimator, we first estimate ATE and its asymptotic variance on each data set and then use sample size weighting to combine the estimated ATE and variances together:

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

(13)

where $\hat{\tau}_{\text{aipw}}^{(k)}$ is the estimated ATE on data set $k$, and $\hat{V}_{\tau}^{(k)}$ is the estimated variance of $\hat{\tau}_{\text{aipw}}^{(k)}$. 

...
This federated AIPW estimator is quite general. First, it is robust to propensity or outcome model misspecification. Second, it allows the propensity or/and outcome models to vary arbitrarily across data sets. Third, it allows $\hat{\tau}_k$ to be estimated from flexible machine learning methods, such as random forests, as we do not need an approach to federate estimated propensity and outcome models across data sets. The tradeoff is that the unrestricted estimator is less efficient than the restricted estimator, under stable covariance distribution and stable propensity and outcome models.

3.4 Practical guidance

In this subsection, we suggest some diagnostic tests that may help practitioners choose between restricted and unrestricted methods and determine the set of shared parameters. For ease of discussion, our empirical application is used as a running example with a generalized linear model (GLM) specification for outcomes.

First, we can examine whether the link function of the GLM is the same across data sets. If not (for example, one is linear and the other one is logit), then it is natural to choose the unrestricted method without shared parameters.

Suppose the link function of the GLM is the same across data sets. Second, we can examine whether there exist some covariates that are unique to a data set. If yes, then it is natural to specify the parameters of these covariates as unstable parameters. For example, Optum covers more years than MarketScan, and the outcome model incorporates several year dummies that are unique to Optum. The coefficients of these dummies are unstable parameters.

Third, we can run hypothesis tests for whether the parameter values are the same across data sets. Suppose we would like to test whether the $p$-dimensional parameters on MarketScan $\beta_M$ and on Optum $\beta_O$ are the same, that is,

$$H_0 : \beta_M = \beta_O \quad H_1 : \beta_M \neq \beta_O.$$  (14)

We can construct the modified Hotelling’s T-square test statistic,

$$T^2 = (\hat{\beta}_M - \hat{\beta}_O)^T \left( \frac{n_M}{(n_M + n_O)^2} \hat{V}_M + \frac{n_O}{(n_M + n_O)^2} \hat{V}_O \right)^{-1} (\hat{\beta}_M - \hat{\beta}_O),$$

where $\hat{\beta}_M$ and $\hat{\beta}_O$ are estimated parameters on MarketScan and on Optum, with estimated asymptotic variances $\hat{V}_M$ and $\hat{V}_O$.

$T^2$ is approximately chi-square distributed with $p$ degree of freedom when both $\hat{\beta}_M$ and $\hat{\beta}_O$ are asymptotically normal. If we do not reject the null, then we can treat $\beta_M$ and $\beta_O$ as stable parameters. Otherwise, we have two options. First, we can treat every entry in $\beta_M$ and $\beta_O$ as an unstable parameter. Second, we can test again on a subset of $\beta_M$ and $\beta_O$ using a similar procedure to determine whether this subset of parameters are stable. We may want to choose the second option when we want to specify as many stable parameters as possible for efficiency consideration (following the intuition in Proposition 1).

Last but not least, we suggest running a data-driven simulation study using real data to compare various federated methods with different specifications of shared and dataset-specific parameters. See Section 5.1 for an example. In this simulation study, we draw patient records from one data set to construct subsamples that mimic the demographics of the multiple data sets we seek to federate. Then we federate subsamples using various federated methods. The benchmarks are the results from the combined data, as in this case, combining patient records across subsamples is permissible, given that they are sampled from one data set. Finally, we choose the federated method that is closest to the benchmarks.

4 ASYMPTOTIC RESULTS

In this section, we show the asymptotic results of our federated MLE, IPW-MLE, and AIPW. The federated point estimators have the same asymptotic distributions as their corresponding estimators using the combined, individual-level data. The federated variance estimators are consistent, which allows us to construct valid confidence intervals of target parameters. Appendix C demonstrate the finite-sample properties of the asymptotic results. Appendix D collects all the proofs.

To show the asymptotic results, we impose standard regularity assumptions on $f(y|x,w,\beta)$ and $e(x,\gamma)$, similar to White and Wooldridge among others. To conserve space, the regularity assumptions are deferred to Assumption
1 in Appendix A.1. Let \( \gamma^{cb} \) and \( \beta^{cb} \) be the solutions that maximize the expected log-likelihood \( \mathbb{E}[\log e^{cb}(x, \gamma)] \) and \( \mathbb{E}[\log f^{cb}(y|x, w, \beta)] \). The solutions may or may not equal the true parameter values \( \gamma^{cb}_0 \) and \( \beta^{cb}_0 \), depending on whether the propensity and outcome models are correctly specified. See Appendix A.1 for more discussion. In this section, we show that our federated MLE or IPW-MLE can consistently estimate \( \beta^{cb} \) (and \( \gamma^{cb} \)).

### 4.1 Federated MLE

We illustrate the asymptotic results of federated MLE using the estimated parameters in the outcome model, but the asymptotic results also apply to the estimated parameters in the propensity model. The following theorem shows that in federated MLE, the federated point estimator of target parameters, denoted by \( \hat{\beta}_{\text{mle}} \), have the same asymptotic distribution as MLE on the combined, individual-level data. In addition, the federated variance estimator, denoted by \( V^{\text{fed}}_{\beta} \), is consistent.

**Theorem 1** (Federated MLE). Suppose Assumption 1.1 holds. If Condition 5 holds, we use restricted federated MLE in Section 3.1.1; otherwise, we use unrestricted federated MLE in Section 3.1.2. Suppose the information matrices satisfy \( \| I^{cb}(\beta)^{-1} I^{(k)}(\beta) \|_2 \leq M \) for some \( M < \infty \) and for all \( k \). As \( n_1, \ldots, n_D \to \infty \), we have

\[
n^{1/2}_{\text{pool}} (\hat{\beta}^{cb}_{\text{mle}} - \beta^{cb}) \xrightarrow{d} \mathcal{N}(0, I_d),
\]

where \( d \) is the dimension of \( \beta^{cb} \). If we replace \( \hat{\beta}^{cb}_{\text{mle}} \) by \( \hat{\beta}^{\text{fed}}_{\beta} \) and/or replace \( \hat{\beta}^{\text{fed}}_{\beta} \) by \( \hat{\beta}^{\text{fed}}_{\beta} \), then (15) continues to hold.

The federated point estimator \( \hat{\beta}^{\text{fed}}_{\beta} \) converges at the optimal rate \( n^{-1/2}_{\text{pool}} \) convergence rate. The convergence rate is therefore improved via federation, as compared to the rate \( n^{-1/2}_k \) of \( \hat{\beta}^{(k)}_{\text{mle}} \) for any \( k \). Theorem 1 holds regardless of whether the outcome model is correctly specified or not. If the outcome model is correctly specified, \( \hat{\beta}^{\text{fed}}_{\beta} \) is a consistent estimator of \( \beta^{cb}_0 \); otherwise, \( \hat{\beta}^{\text{fed}}_{\beta} \) converges to the limit \( \beta^{cb} \) that generally differs from \( \beta^{cb}_0 \).

**Remark 1.** If outcome models are unstable, but we use restricted federated MLE in Section 3.1.1. Proposition 2 in Appendix A shows that, under some special cases, Theorem 1 continues to hold, but with a limit that potentially differs from \( \beta^{cb}_0 \).

### 4.2 Federated IPW-MLE

We start with a lemma that provides the asymptotic distribution of IPW-MLE on a single data set, on which the asymptotic results of federated IPW-MLE are built.

**Lemma 1.** Suppose Assumption 1 holds and we estimate \( e(X_i) \) from MLE. As \( n \to \infty \), \( \hat{\beta}_{\text{ipw-mle}} \) estimated from IPW-MLE is consistent and asymptotically normal,

\[
\sqrt{n} \left( \hat{\beta}_{\text{ipw-mle}} - \beta^* \right) \xrightarrow{d} \mathcal{N} \left( 0, V_{\beta^*, \text{ipw-mle}, \ell} \right),
\]

where

\[
V_{\beta^*, \text{ipw-mle}, \ell} = A_{\beta^*, \ell}^{-1} (D_{\beta^*, \ell} - M_{\beta^*, \ell}) A_{\beta^*, \ell}^{-1},
\]

with

\[
M_{\beta^*, \ell} = \begin{cases} 
C_{\beta^*, \ell} V_\ell C_{\beta^*, \ell}^\top & \text{ATE weighting} \\
C_{\beta^*, \ell} V_\ell C_{\beta^*, \ell, 2}^\top + C_{\beta^*, \ell, 2} V_\ell C_{\beta^*, \ell, 1}^\top - C_{\beta^*, \ell, 2} V_\ell C_{\beta^*, \ell, 2}^\top & \text{ATT weighting}
\end{cases}
\]
\[ V_{\beta} = A_{\beta}^{-1} B_{\beta} A_{\beta}^{-1} \]

and \( A_{\beta,m} \) is matrix \( A_{\beta,m} \) evaluated at \( \beta^* \), with the definition of \( A_{\beta,m} \) provided in Table 1.

Other terms in formula (16) are defined similarly.

If IPW-MLE uses true propensities, then the asymptotic variance is simplified to

\[ V_{\beta,\text{ipw-mle}} = A_{\beta,m}^{-1} D_{\rho,m} A_{\beta,m}^{-1}. \]  \( \text{(17)} \)

Lemma 1 coincides with the results in Wooldridge for ATT weighting, and Lemma 1 additionally provides the results for ATT weighting.

Note that the estimation error of the propensity model carries over to the asymptotic variance of IPW-MLE. This explains why our federated variance estimator in Section 3.2 needs to vary with whether the true propensities are used. In addition, if federated IPW-MLE varies properly with whether propensity and/or outcome models are stable or not, then the federated point estimator, denoted by \( \hat{\beta}_{\text{ipw-mle}} \), have the same asymptotic distribution as IPW-MLE on the combined, individual-level data. Moreover, the federated variance, denoted by \( V_{\beta,\text{fed},\epsilon} \), is consistent when it is obtained based on the formulas in Lemma 1.

**Theorem 2** (Federated IPW-MLE). Suppose Assumption 1 holds. If Conditions 4 and 5 hold, we use restricted federated IPW-MLE in Section 3.2.1; otherwise, we use unrestricted federated IPW-MLE in Section 3.2.2. Suppose \( \left\| (A_{\beta,m}^{cb})^{-1} A_{\beta,m}^{cb} \right\| \leq M \) and \( \left\| (A_{\beta,m}^{cb})^{-1} A_{\beta,m}^{cb} \right\| \leq M \) for some \( M < \infty \) and for all \( k \). As \( n_1, \ldots, n_D \to \infty \), we have

\[ n_{\text{pool}}^{1/2} (V_{\beta,\text{ipw-mle}})^{-1/2} \left( \hat{\beta}_{\text{ipw-mle}} - \beta^{cb} \right) \xrightarrow{d} \mathcal{N}(0, I_4). \]  \( \text{(18)} \)

If we replace \( V_{\beta,\text{ipw-mle}} \) by \( V_{\beta,\text{ipw-mle},\epsilon} \) and/or replace \( \hat{\beta}_{\text{ipw-mle}} \) by \( \hat{\beta}_{\text{ipw-mle}}^{cb} \), then \( \text{(18)} \) continues to hold. If we use true propensities, then \( \text{(18)} \) continues to hold with \( V_{\beta,\text{ipw-mle},\epsilon} \) replaced by the corresponding variance terms for the true propensities.

Federated IPW-MLE converges at the rate \( n_{\text{pool}}^{-1/2} \), which is faster than \( n_k^{-1/2} \) on a single data set \( k \). This theorem holds regardless of whether covariate distributions are stable or not, as long as the limiting objects, \( \beta^{cb} \) and \( V_{\beta,\text{ipw-mle}}^{cb} \), are well-defined on the combined data, though their definitions may vary with whether covariate distributions are stable.

Moreover, Theorem 2 holds regardless of whether we use the true or estimated propensities. In practice, for ATE weighting, even if we know the true propensities, it is better to use the estimated propensities for the efficiency consideration as \( V_{\beta,\text{ipw-mle}} - V_{\beta,\text{ipw-mle},\epsilon} \) is positive semidefinite from Lemma 1. If the estimated propensities are used, we could still use the federated variance estimator for the true propensity case, which takes a simpler form, but at the cost of overestimating the variance.

### 4.3 Federated AIPW

The following theorem shows that our federated AIPW for ATE and ATT has the same asymptotic distribution as AIPW on the combined data. In addition, our federated variance estimators for AIPW are consistent.

**Theorem 3.** Suppose either of the following cases holds: (a) the score \( d^{(k)}(x, w, y) \) is the same for all \( k \), and we use the federation procedure in Section 3.3.1; or (b) \( d^{(k)}(x, w, y) \) varies with \( k \), and we use the federation procedure in Section 3.3.2. Furthermore, suppose for any data set \( k \), at least one condition holds: (a) \( \mu^{(k)}(x) \) is correctly specified and consistently estimated for \( w \in \{0, 1\} \), or (b) \( e^{(k)}(x) \) is correctly specified and consistently estimated. As \( n_1, \ldots, n_D \to \infty \), if the estimand is ATE, we have

\[ n_{\text{pool}}^{1/2} (V_{\tau_{ae}})^{-1/2} (\hat{\tau}_{\text{fed}} - \tau_{\text{ate}}) \xrightarrow{d} \mathcal{N}(0, 1). \]  \( \text{(19)} \)

If we replace \( V_{\tau_{ae}} \) by \( V_{\tau_{me}} \) and/or replace \( \hat{\tau}_{\text{fed}} \) by \( \hat{\tau}_{\text{ate}} \), then \( \text{(19)} \) continues to hold. If the estimand is ATT, \( \text{(19)} \) continues to hold analogously for the federated estimator of ATT and the corresponding federated variance estimator.
Analogous to federated MLE and IPW-MLE, federated AIPW achieves a faster convergence rate than AIPW on a single data set. The estimation efficiency of ATE and ATT can be improved through federation. In addition, federated AIPW achieves the semiparametric efficiency bound.

Note that if the propensity and outcome models are estimated from flexible machine learning methods, then we can use unrestricted federated AIPW to combine the estimated ATE or ATT on individual data sets together without combining individual propensity and outcome models. If individual propensity and outcome models can be estimated at rate \( o(n^{-1/4}) \), then the estimated ATE or ATT on individual data sets by using cross-fitting converges at the rate \( n^{1/2} \) and is asymptotically normal.48 We can then show that the federated ATE or ATT is asymptotically normal, and the federated variance estimator is consistent. This approach is asymptotically efficient. However, when propensity and outcome models are stable, the variance may be reduced in finite samples, by developing new approaches to federate flexible machine learning methods.

5 | EMPIRICAL STUDIES BASED ON MEDICAL CLAIMS DATA

In this section, we further study the effect of alpha blockers on two distributed medical databases, MarketScan and Optum, introduced in Section 1.‡‡‡ We first evaluate various federation methods through a data-driven simulation study on one medical claims data, select the optimal federated method, and then apply this method to federate MarketScan and Optum.§§§

5.1 | Simulation on one medical claims data set

In the data-driven simulation study, we first construct subsamples from one cohort to reflect patient demographics from MarketScan and Optum. Next we compare estimates from various federated methods with those on the combined data. We seek to evaluate 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 5.2.

We start by presenting our approach to simulate subsamples from one patient cohort in Section 5.1.1. Then we list benchmark methods and tested federated methods in Section 5.1.2. We compare the results from federated methods against benchmarks in Section 5.1.3.

5.1.1 | Sampling schemes for subsamples

We draw two subsamples, denoted as \( S_1 \) and \( S_2 \), based on patient records from one cohort in a database (denoted as \( C \)), to mimic the demographics of the distributed databases that we aim to federate. Our simulation design is based on the observation that cohorts in MarketScan include patients younger than age 65 from 2009 to 2015, while cohorts in Optum include patients up to age 85 from 2005 to 2019, with a majority to be over age 65.

To simulate subsamples, we first partition one cohort \( C \) into four disjoint sub-cohorts, denoted as \( C_1, C_2, C_3, \) and \( C_4 \), by age and fiscal year. \( C_1 \) include patients younger than the median age of \( C \) up to year 2012; \( C_2 \) include patients younger than the median age after 2012; \( C_3 \) include patients older than the median age up to 2012; \( C_4 \) include patients older than the median age after 2012. Next we simulate \( S_1 \) and \( S_2 \). \( S_1 \) mimics the demographics of MarketScan, with 70% and 30% sampled from \( C_1 \) and \( C_3 \), respectively, with replacement. \( S_2 \) mimics the demographics of Optum, with 10%, 10%, 10%, and 70% are sampled from \( C_1, C_2, C_3, \) and \( C_4 \), respectively, with replacement.

As a robustness check, we consider other approaches in Appendix B to construct subsamples, including varying the sampling ratios from different sub-cohorts, varying subsample sizes, and varying the number of subsamples.

5.1.2 | Estimation and benchmarks

We consider two benchmark estimators and three federated estimators.
Restricted benchmarks
Parameters in the propensity and outcome models are assumed to be stable across subsamples. On the combined data, we specify restricted propensity and outcome models as

\[
\frac{\text{pr}(W_i = 1|X_i)}{\text{pr}(W_i = 0|X_i)} = X_i^T \gamma \\
\frac{\text{pr}(Y_i = 1|X_i, W_i)}{\text{pr}(Y_i = 0|X_i, W_i)} = W_i \beta_w + X_i^T \beta_X.
\] (20)

where outcome \(Y_i\) is binary indicating whether a patient received mechanical ventilation and then had an in-hospital death \((Y_i = 1)\) or not \((Y_i = 0)\), treatment \(W_i\) is binary indicating whether a patient is exposed to alpha blockers \((W_i = 1)\) or not \((W_i = 0)\), and \(X_i\) consists of age, fiscal year dummies, and health-related confounders.

The restricted benchmarks are the estimate of \(\beta_w\) in (20), denoted as \(\hat{\beta}_{w, bm}\), and its estimated variance, denoted as \(\hat{V}_{\beta_{w, bm}}\), from the combined data.

Unrestricted benchmarks
Parameters in the propensity and outcome models can be unstable across subsamples. On the combined data, we specify a flexible functional form for the propensity and outcome models

\[
\frac{\text{pr}(W_i = 1|X_i)}{\text{pr}(W_i = 0|X_i)} = X_i^T \gamma_s + X_i^T \gamma_{uns} \left(1(A_i = 1) \gamma_{uns}^{(1)} + 1(A_i = 2) \gamma_{uns}^{(2)}\right) \\
\frac{\text{pr}(Y_i = 1|X_i, W_i)}{\text{pr}(Y_i = 0|X_i, W_i)} = W_i \beta_w + X_i^T \beta_X s + X_i^T \beta_{X, uns} \left(1(A_i = 1) \beta_{X, uns}^{(1)} + 1(A_i = 2) \beta_{X, uns}^{(2)}\right),
\] (21)

where \(A_i \in \{1, 2\}\) indicates whether the patient record belongs to \(S_1\) or \(S_2\). Parameters are partitioned into stable parameters \((\gamma_s, \beta_w, \beta_X s)\) and unstable parameters \((\gamma_{uns}, \beta_{X, uns}^{(a)}\) for \(a \in \{1, 2\}\)). The unstable variables include the coefficients of age confounders and year dummies unique to \(S_a\), which is motivated by the observation that age coefficient has opposite signs on MarketScan and Optum (see Figure B.4 in Appendix B), and Optum covers more years. Note that \(\beta_w\) is stable across subsamples, which can be interpreted as the average treatment coefficient across subsamples.

The unrestricted benchmarks are the estimates of \(\beta_w\) in (21), denoted as \(\hat{\beta}_{w, bm}\), and its estimated variance, denoted as \(\hat{V}_{\beta_{w, bm}}\), from the combined data.

Restricted federated estimators
Under the restricted model specification (20), we use restricted federated IPW-MLE to estimate \(\beta_w\) in (20) and its variance. Let \(\hat{\beta}_{w, ipw-mle}^{\text{fed}}\) and \(\hat{V}_{\beta_w, ipw-mle}^{\text{fed}}\) be the estimated coefficient and variance.

Unrestricted federated estimators
Under the flexible model specification (21), we use our unrestricted federated IPW-MLE to estimate \(\beta_w\) in (21) and its variance. Let \(\hat{\beta}_{w, ipw-mle}^{\text{fed}}\) and \(\hat{V}_{\beta_w, ipw-mle}^{\text{fed}}\) be the estimated coefficient and variance.

Inverse variance weighting
Under the restricted model specification (20), we use IVW to estimate \(\beta_{0, w}\) in (20) and its variance. Let \(\hat{\beta}_{w, ivw}\) and \(\hat{V}_{\beta_{0, w}, ivw}\) be the estimated coefficient and variance.

5.1.3 Results
We compare restricted and unrestricted federated IPW-MLE and IVW with the restricted and unrestricted benchmarks in Table 5. Additional simulation results with alternative sampling schemes and with federated MLE are presented in Tables B2-B4 in Appendix B.5. The error of a federated estimator is defined as its difference from the benchmark.

There are four observations from Table 5. First and foremost, for both point and variance estimates, our restricted and unrestricted federated IPW-MLE have much lower errors than IVW, when compared to restricted and unrestricted benchmarks. Second, the restricted federated point estimator is closer to the restricted benchmark than the unrestricted
TABLE 5 Comparison between restricted/unrestricted federated estimators and IVW with corresponding restricted/unrestricted benchmarks.

| Restricted benchmarks $\hat{\beta}_{w,bm}^r$, $\hat{\gamma}_{w,bm}^r$ | $\hat{\beta}_{w,ivw}$ MAE | $\hat{\gamma}_{w,ipw-mle}$ MAE | $\hat{\beta}_{w,ipw-mle}$ MAE |
|-------------------------|-----------------|-----------------|-----------------|
| ARD | -0.6757 | 1.2349 | 0.0538 | 0.0677 |
| PNA | -0.3250 | 0.6482 | 0.0541 | 0.0384 |

| Unrestricted benchmarks $\hat{\beta}_{w,bm}$, $\hat{\gamma}_{w,bm}$ | $\hat{\beta}_{w,ivw}$ MAE | $\hat{\gamma}_{w,ipw-mle}$ MAE | $\hat{\beta}_{w,ipw-mle}$ MAE |
|-------------------------|-----------------|-----------------|-----------------|
| ARD | 0.1098 | 0.0848 | 0.0395 | 0.0363 |
| PNA | 0.0641 | 0.0376 | 0.0158 | 0.0129 |

Note: Subsamples are simulated from the MarketScan ARD cohort, and from the MarketScan pneumonia (PNA) cohort with $D = 2$. For subsamples drawn from ARD cohort, $n_1 = n_2 = 6000$; for subsamples drawn from PNA cohort, $n_1 = n_2 = 10000$. We use ATE weighting in IPW-MLE in these tables. The mean absolute error (MAE) is calculated relative to the benchmark mean values (first column of each table) based on 50 iterations of independent draws of subsamples. We report the mean value of benchmarks because the combined data $C_1 \cup C_2$ from which benchmarks are estimated vary across iterations.

benchmark. Analogously, the unrestricted federated point estimator is closer to the unrestricted benchmark. Third, the variance in the unrestricted benchmark and federated variance are larger than the restricted counterparts, implying the efficiency loss when flexible model specifications are used. Fourth, interestingly, the unrestricted federated variance is closer to variances in both restricted and unrestricted benchmarks. This is because federated variance tends to underestimate the true variance in finite samples (even though both are consistent). As the unrestricted federated variance tends to be larger, it partially corrects for the underestimation error.

These observations are robust to alternative sampling schemes and to federated MLE as shown in Tables B2-B4 in Appendix B.5.‡‡‡‡ As unrestricted federated IPW-MLE is more flexible and generally provides a better variance estimate, we use unrestricted federated IPW-MLE to federate MarketScan and Optum, as shown in Section 5.2 below.

5.2  | Federation across two medical claim data sets

In this section, we seek to federate MarketScan and Optum to study the effect of alpha-blockers. As shown in Figure 5, the coefficient on alpha blockers is consistently negative on the individual cohorts of ARD patients and of pneumonia patients, implying a reduced risk of adverse outcomes for ARD and pneumonia patients who were exposed to alpha blockers.

However, coefficients of some confounders, for example, age, are of different magnitudes or signs in the outcome model across the two databases (though, none with statistical significance). 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 B1, as well as the heterogeneous response of outcomes to the treatment and confounders. Unrestricted federated IPW-MLE with a flexible functional form for the combined data seems to be preferable 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). Third, we have largely controlled for unobserved confounders in our approach to constructing cohorts, as discussed in Appendix B.3, and sensitivity analyses are conducted in Koenecke et al.1
**Figure 5** Federation across MarketScan and Optum. These figures show the estimated coefficient of alpha blockers and 95% confidence interval on MarketScan and Optum, and federated coefficient and 95% confidence interval from unrestricted federated IPW-MLE with ATE weighting. Note that for the pneumonia cohort, the confidence intervals for the federated estimator are wider than those on Optum. This can happen when the asymptotic variance is heterogeneous across data sets and the asymptotic variance on the small data is much larger than on the large data. In this case, the federated variance obtained by sample size weighting can be larger than the variance on the large data. See Appendix B.2 for a toy example. See Figure B3 in Appendix B for ATT weighting; the results are close to those in these figures.

Figure 5 shows the federated point estimates and confidence intervals from unrestricted federated IPW-MLE. 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 Figures 1 and B3 in Appendix B).

As a robustness check, we report the results from federated MLE in Figure B2, and estimated treatment effects from federated IPW-MLE and AIPW in Figure B5 in Appendix B. Both the coefficient in the outcome model and estimated treatment effects of alpha blockers are negative and statistically significant, supporting our finding of an association between exposure to alpha blockers and risk reduction of progression to ventilation and death.

### 6 CONCLUSION

This article proposes three categories of federated inference methods based on MLE, IPW-MLE, and AIPW, respectively. Our federated point estimators have the same asymptotic distributions as the corresponding estimators from combined, individual-level data. Our federated variance estimators are consistent. To achieve these properties, we show that the implementations of our federated methods should be adjusted based on conditions such as whether propensity and outcome models are stable across heterogeneous data sets. Finally, we apply our federated inference methods to study the effectiveness of alpha blockers on patient outcomes from two separate medical claims databases.

To conclude, we would like to point out three interesting directions for future work. The first is to develop federated semiparametric or nonparametric estimation methods. The second is to develop communication-efficient, theoretically guaranteed federated causal inference methods in settings with high-dimensional nuisance parameters. The third is to develop these methods in settings with many data sets, while each data set may only have a small number of observations.

### FUNDING INFORMATION

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 thank Kristine Koutout, Molly Offer-Westort, seminar and conference participants at Berkeley, Microsoft Research, and American Causal Inference Conference for helpful comments.

### DATA AVAILABILITY STATEMENT

Data for these analyses were made available to the authors through the third-party license from Optum and Truven (two commercial data providers in the United States). As such, the authors cannot make these data publicly available due to the data use agreement. Other researchers can access these data by purchasing a license through Optum and Truven. Inclusion criteria can be found in the “Materials and methods” section (specifically, the “Study Definition” subsection) in Koenecke et al. that would allow other researchers to identify the same cohort of patients we used for the empirical applications. Interested individuals may see https://www.optum.com/solutions/prod-nav/product-data.html for more information on accessing Optum data, https://marketscan.truvenhealth.com/marketscanportal/ for more information on accessing Truven (MarketScan) data. Code is available at https://github.com/ruoxuanxiong/federated-causal-inference. Data for this project were accessed using the Stanford Center for Population Health Sciences Data Core.
ENDNOTES

* Treated group that is exposed to alpha blockers has more elderly patients than control groups. This is because alpha blockers are commonly prescribed for chronic prostatitis, and the prostate generally worsens with age.
† IPW-MLE includes linear models, logit models, Poisson models, and Cox models weighted by inverse propensity scores as special cases.
‡ Overlooking this effect can lead to an overestimate of variance and a loss of efficiency.
§ IVW is asymptotically the same as our federated IPW-MLE when data sets are homogeneous in the sense that covariate distributions, as well as propensity and outcome models, are stable across data sets.
¶ IPW logistic regression is a special case of IPW-MLE.

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. See Appendix B.1 for a numerical example for more intuition.

Early developments in data mining provide methods to combine point estimates of model parameters in linear models, logit models, and maximum likelihood estimators across distributed information systems, with most methods being iterative. Recent advances, mainly in federated learning, aim to develop communication-efficient methods to optimize parameters across a large number of distributed heterogeneous agents, while preserving privacy. Importantly, statistical inference is not a primary consideration in the aforementioned literature.

An iterative approach can provide estimators that are closer to those from the pooled individual-level data. However, we show that the difference between iterative and non-iterative approaches can be neglected asymptotically.

There has been a growing literature surrounding the development of causal inference methods, when individual-level data can be shared across multiple data sets, but data sets are collected under heterogeneous conditions. The parameters in \( \hat{\mu}_{ij}(x) \) and \( \hat{e}(x) \) are omitted to account for the case where \( \hat{\mu}_{ij}(x) \) and \( \hat{e}(x) \) are estimated by nonparametric methods when the individual-level data could have been combined.

For ease of presentation, 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. If there are some shared parameters across several but not all data sets, we just need to combine these parameters in \( \beta^b \). For example, if \( \beta_{uns}^j \) and \( \beta_{uns}^k \) are the same for \( j \) and \( k \), then we merge \( \beta_{uns}^j \) and \( \beta_{uns}^k \) in \( \beta^b \).

Age coefficient has opposite signs in the two data sets in our empirical study, as shown in Figure B.4.

If the outcome model is correctly specified, the information matrix equivalence holds, implying that \( A^e = A^e_\beta \) and \( V^e = A^{-1}_\beta \). Then we only need to estimate and combine \( A^e_\beta \).

Zero-padding is a commonly used technique in signal processing and deep learning to preprocess inputs to the same length.

When the true propensity model is known and used, we do not need to federate the individual propensity models.

The unrestricted AIPW is equivalent to the AIPW in (4) with the score on combined data estimated by \( \hat{\phi}(X_i^{(k)}, W_i^{(k)}, Y_i^{(k)}) = \prod_{j=1}^K \hat{\phi}(X_i^{(j)}, W_i^{(j)}, Y_i^{(j)}) \) and \( \hat{\phi}(X_i^{(j)}, W_i^{(j)}, Y_i^{(j)}) \) estimated using the estimated outcome and propensity models on data set \( j \).

Our analysis builds on the studies by König et al., Koencke et al., Rose et al., and Powell et al.

Note that our findings reproduce similar results to Koencke et al., validating the prior result suggesting that alpha blockers are effective in reducing ventilation and death in ARD and pneumonia patients; however, the confidence levels are narrower because our federated methods presented here are improved from those used in Koencke et al. Koencke et al. only use the treatment coefficient and variance in federation, whereas here, we use the full variance-covariance matrix from all covariates. Our approach leverages the stable part of the model across two data sets, which could improve the estimation precision of the treatment coefficient.

See Appendix B.3 for the full list of confounders.

Equation (21) can be easily generalized to the case with more than two subsamples.

Note that \( S_1 \) has patient records up to 2012, while \( S_2 \) has patient records for all years. The coefficients of year dummies after 2012 are treated as unstable parameters in both restricted and unrestricted benchmarks.

Age confounders include age, age-squared, and age-cubed.

IVW is appropriate when Conditions 4–6 hold. In this case, Hessians and other matrices in the asymptotic variance are asymptotically stable across data sets. Then we can show that our federated estimators in Section 3 are asymptotically the same as IVW.

We could use alternative approaches to obtaining federated maximum likelihood estimator of treatment coefficient, such as by using a surrogate likelihood function that communicates gradients only or that communicates both gradients and Hessians similar to our federated MLE. In the likelihood function, the heterogeneity in data sets can be adjusted through tilting the density ratio; moreover, a regularization term can be included in high-dimensional settings. These methods do not account for the treatment selection bias and are iterative, while federated IPW-MLE does and is noniterative. We expect the results of these methods to be conceptually similar to those of federated MLE.

Similar to Footnote 24, we could use alternative approaches to obtaining the federated estimator of treatment coefficient. The results would be conceptually similar to those of federated MLE in Figure B2. Due to the treatment selection bias, the estimated treatment coefficient from alternative approaches would not have the interpretation of the average treatment coefficient on either the whole population or the treated population, while the estimated coefficient from IPW-MLE does.

ORCID

Ruoxuan Xiong https://orcid.org/0000-0002-3701-4428
Zhu Shen https://orcid.org/0000-0003-4564-5438
REFERENCES

1. Koenecke A, Powell M, Xiong R, et al. Alpha-1 adrenergic receptor antagonists to prevent hyperinflammation and death from lower respiratory tract infection. *Elife*. 2021;10:e61700.

2. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005;61(4):962-973.

3. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. *J Econom*. 2007;141(2):1281-1301.

4. Wooldridge JM. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Port Econ J*. 2002;1(2):117-139.

5. Du W, Han YS, Chen S. Privacy-preserving multivariate statistical analysis: linear regression and classification. *Proceedings of the 2004 SIAM International Conference on Data Mining*. Philadelphia, PA: SIAM; 2004:222-233.

6. Karr AF, Lin X, Sanil AP, Reiter JP. Secure regression on distributed databases. *J Comput Graph Stat*. 2005;14(2):263-279.

7. Fienberg SE, Fulp WJ, Slavkovic AB, Wrobel TA. "Secure" log-linear and logistic regression analysis of distributed databases. In: Domingo-Ferrer J, Franconi L, eds. *International Conference on Privacy in Statistical Databases*. Berlin: Springer; 2006:277-290.

8. Slavkovic AB, Nardi Y, Tibbits MM. "Secure" logistic regression of horizontally and vertically partitioned distributed databases. *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*. New York: IEEE; 2007:723-728.

9. Blatt D, Hero A. Distributed maximum likelihood estimation for sensor networks. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. New York: IEEE; 2004:iii-929.

10. Karr AF, Fulp WJ, Vera F, Young SS, Lin X, Reiter JP. Secure, privacy-preserving analysis of distributed databases. *Dent Tech*. 2007;49(3):335-345.

11. Zhao T, Nehorai A. Information-driven distributed maximum likelihood estimation based on Gauss-Newton method in wireless sensor networks. *IEEE Trans Signal Process*. 2007;55(9):4669-4682.

12. Lin X, Karr AF. Privacy-preserving maximum likelihood estimation for distributed data. *J Priv Confid*. 2010;1(2):213-222.

13. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492, 2016.

14. McMahan B, Moore E, Ramage D, Hampson S, yBA A. Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics*. Cambridge, MA: PMLR; 2017:1273-1282.

15. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag*. 2020;37(3):50-60.

16. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.

17. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med*. 1991;10(11):1665-1677.

18. Toh S, Wellman R, Coley RY, et al. Combining distributed regression and propensity scores: a doubly privacy-protecting analytic method for multicenter research. *Clin Epidemiol*. 2018;10:1773-1786.

19. Toh S, Rifas-Shiman SL, Lin PID, et al. Privacy-protecting multivariable-adjusted distributed regression analysis for multi-center pediatric study. *Pediatr Res*. 2020;87(6):1086-1092.

20. Duan R, Boland MR, Liu Z, et al. Learning from electronic health records across multiple sites: a communication-efficient and privacy-preserving distributed algorithm. *J Am Med Inform Assoc*. 2020;27(3):376-385.

21. Shu D, Young JG, Shu S. Privacy-protecting estimation of adjusted risk ratios using modified Poisson regression in multi-center studies. *BMC Med Res Methodol*. 2019;19(1):1-7.

22. Shu D, Yoshida K, Fireman BH, Toh S. Inverse probability weighted cox model in multi-site studies without sharing individual-level data. *Stat Methods Med Res*. 2020;29(6):1668-1681.

23. Shu D, Young JG, Toh S, Wang R. Variance estimation in inverse probability weighted cox models. *Biometrics*. 2020;77:1101-1117.

24. Wolfson M, Wallace SE, MascaN, et al. DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol*. 2010;39(5):1372-1382.

25. Jordan M, Lee JD, Yang Y. Communication-efficient distributed statistical inference. *J Am Stat Assoc*. 2018;114:668-681.

26. Duan R, Ning Y, Chen Y. Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika*. 2022;109(1):67-83.

27. Vo TV, Hoang TN, Lee Y, Leong TY. Federated estimation of causal effects from observational data. arXiv preprint arXiv:2106.00456, 2021.

28. Han L, Hou J, Cho K, Duan R, Cai T. Federated adaptive causal estimation (FACE) of target treatment effects. arXiv preprint arXiv:2112.09313, 2021.

29. Han L, Li Y, Niknam BA, Zubizarreta JR. Privacy-preserving and communication-efficient causal inference for hospital quality measurement. arXiv preprint arXiv:2203.00768, 2022.

30. Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. *J R Stat Soc Series B Stat Methodology*. 2016;78:947-1012.

31. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci*. 2016;113(27):7345-7352.

32. Rosenman E, Owen AB, Baiocchi M, Banack H. Propensity score methods for merging observational and experimental datasets. arXiv preprint arXiv:1804.07863, 2018.

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

34. Athey S, Chetty R, Imbens G. Combining experimental and observational data to estimate treatment effects on long term outcomes. arXiv preprint arXiv:2006.09676, 2020.
35. Rothenhäusler D, Meinshausen N, Bühlmann P, Peters J. Anchor regression: heterogeneous data meet causality. *J R Stat Soc Series B Stat Methodology*. 2021;83(2):215-246.
36. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press; 2015.
37. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
38. Speran
del S. Understanding logistic regression analysis. *Biochem Med*. 2014;24(1):12-18.
39. Singh R, Mukhopadhyay K. Survival analysis in clinical trials: basics and must know areas. *Perspect Clin Res*. 2011;2(4):145-148.
40. Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. *J Health Econ*. 1999;18(2):153-171.
41. Blough DK, Ramsey SD. Using generalized linear models to assess medical care costs. *Health Serv Outcomes Res Methodol*. 2000;1(2):185-202.
42. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866.
43. White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982;50:1-25.
44. Madan B, Bein D. Optimal maximum likelihood estimates fusion in distributed network of sensors. *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*. New York: IEEE; 2016:369-375.
45. O’Shea K, Nash R. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458, 2015.
46. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2018;113(523):1228-1242.
47. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71(4):1161-1189.
48. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. Double/debiased/Neyman machine learning of treatment effects. *Am Econ Rev*. 2017;107(5):261-265.
49. Konig MF, Powell M, Staedtke V, et al. Preventing cytokine storm syndrome in COVID-19 using α-1 adrenergic receptor antagonists. *J Clin Invest*. 2020;130(7):3345-3347.
50. Rose L, Graham L, Koenecke A, et al. The association between Alpha-1 adrenergic receptor antagonists and In-hospital mortality from COVID-19. *Front Med*. 2021;8:304. doi:10.3389/fmed.2021.637647
51. Powell M, Koenecke A, Byrd JB, et al. Ten rules for conducting retrospective pharmacoepidemiological analyses: example COVID-19 study. *Front Pharmacol*. 2021;12:1799.
52. Wang J, Kolar M, Srebro N, Zhang T. Efficient distributed learning with sparsity. *International Conference on Machine Learning*. Cambridge, MA: PMLR; 2017:3636-3645.
53. Li S, Cai T, Duan R. Targeting underrepresented populations in precision medicine: a federated transfer learning approach. arXiv preprint arXiv:2108.12112, 2021.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Xiong R, Koenecke A, Powell M, Shen Z, Vogelstein JT, Athey S. Federated causal inference in heterogeneous observational data. *Statistics in Medicine*. 2023;42(24):4418-4439. doi: 10.1002/sim.9868