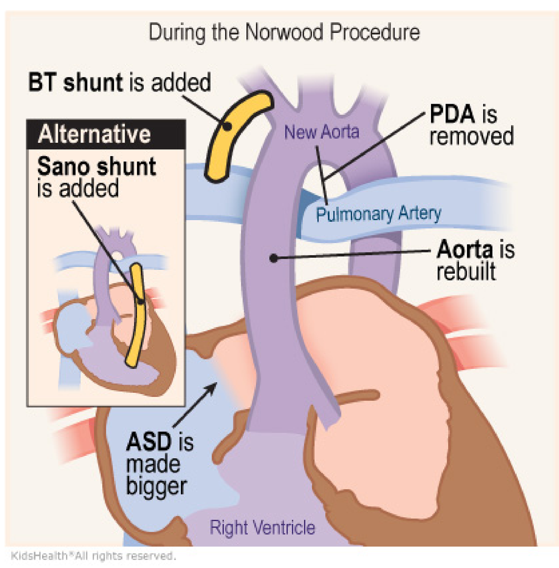
Multi-Source Conformal Inference Under Distribution Shift

Yi Liu[1] Alexander W. Levis[2] Sharon-Lise Normand[3] Larry Han[4]

[1]Department of Statistics, North Carolina State University
[2]Department of Statistics and Data Science, Carnegie Mellon University
[3] Department of Health Care Policy, Harvard Medical School
[4] Bouvé College of Health Sciences, Northeastern University

Scan for the full ICML paper!

Motivating Case Study: Congenital Heart Defects



- **Impact**: Congenital heart defects (CHD) are the most common birth defects in the US.
- **Data Source**: STS-CHSD database. We focused on Norwood surgeries performed from 2016-2022.
- **Outcome**: Post-surgery length of stay (LOS) in hospital.
- **Observations**: There were 3,457 observations with a median LOS of 40 days (min: 2, max: 183), with 752 (21.2%) missing LOS values.
- **Goal**: For a new patient who arrives at the hospital, can we provide a conformal prediction interval[2] $\widehat{C}(\boldsymbol{x})$ that will contain the true LOS with some pre-specified coverage level $1-\alpha$:

$$\mathbb{P}(Y \in \widehat{C}(\boldsymbol{X})) \geq 1-\alpha.$$

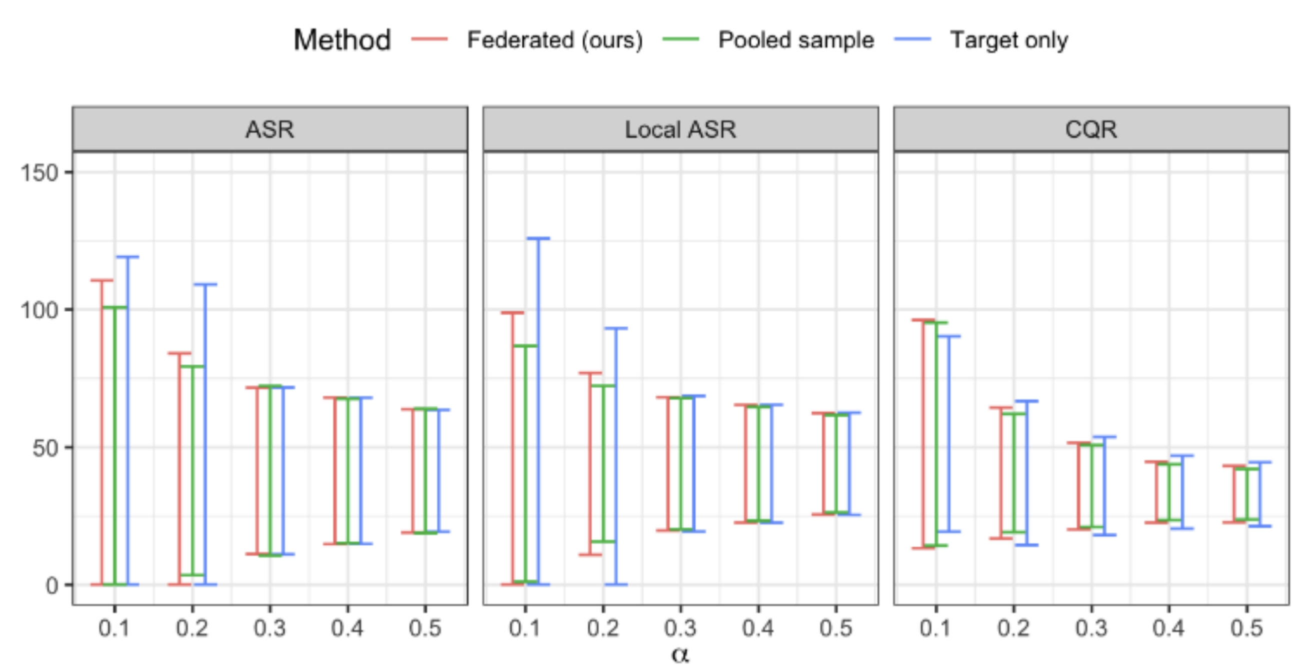


Figure 1. Prediction intervals for hospital LOS for a randomly selected patient across miscoverage levels $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and conformal scores $\in$ {ASR, local ASR, CQR}.

Notation and Set-up

- Data from K sites. Let $T \in \{0, 1, ..., K-1\}$ denote study sites. $T = 0$ indicates the target site, and the rest are source sites.
- R is an indicator for observing outcome Y: $R = 1$ if Y is observed, $R = 0$ if missing.
- Data: random sample of n i.i.d. copies of $\mathcal{O} = (\boldsymbol{X}, T, R, RY) \sim \mathbb{P}$.
- Assumption 1 (**Missing at random [MAR]**). $R \perp Y \mid T, \boldsymbol{X}$.
- Assumption 2 (**Positivity**). For $\epsilon > 0$, $\mathbb{P}(R = 1 \mid T, \boldsymbol{X}) \geq \epsilon$ with probability 1.
- Two important goals of conformal inference:
  - Distribution-free: valid in finite samples for any $(\boldsymbol{X}, Y)$ and any predictive algorithm.
  - Efficient: to minimize width of interval $\widehat{C}(\boldsymbol{X})$.

References

[1] Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *arXiv preprint arXiv:2112.09313, 2021.*

[2] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[3] Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae009, 2024.

Efficient Multi-Source Predictions

- Given the set-up, our goal is to construct prediction intervals $\widehat{C}_\alpha(\boldsymbol{X})$, $\alpha \in (0, 1)$, such that

$$\mathbb{P}(Y \in \widehat{C}_\alpha(\boldsymbol{X}) \mid T = 0, R = 0) \geq 1 - \alpha.$$

- Predictions tailored for missing outcomes in the target site with marginal coverage guarantees.
- Introduce a conformal score, $S(\boldsymbol{X}, Y)$. Predictions: $\widehat{C}_\alpha(\boldsymbol{X}) = \{y \in \mathbb{R} : S(\boldsymbol{X}, y) \leq \widehat{r}\}$.
- $\widehat{r}$ estimates $r_0 = r_0(\alpha)(\mathbb{P})$, the $(1-\alpha)$-quantile of $S(\boldsymbol{X}, Y)$.
- Under MAR, r_0 is identified by the following equation, using target site data only:

$$1 - \alpha = \mathbb{P}(S(\boldsymbol{X}, Y) \leq r_0 \mid T = 0, R = 0) = \mathbb{E}(\mathbb{P}(S(\boldsymbol{X}, Y) \leq r_0 \mid T = 0, \boldsymbol{X}, R = 1) \mid T = 0, R = 0).$$

- **Common Conditional Outcomes Distribution (CCOD) in Multi-Source Data.** If the CCOD holds, we propose the following efficient influence function (IF)[3] of $r_0 = r_0(\alpha)(\mathbb{P})$:

$$I(T = 0)(1 - R)\{\overline{m}(r_0, \boldsymbol{X}) - (1 - \alpha)\} + R\overline{\eta}(\boldsymbol{X})q_0(\boldsymbol{X})\{I(S(\boldsymbol{X}, Y) \leq r_0) - \overline{m}(r_0, \boldsymbol{X})\}$$
$$:= \varphi^{\text{CCOD}}(\mathcal{O}; r_0, \overline{m}, \overline{\eta}, q_0),$$

where
  - $\overline{m}(r, \boldsymbol{X}) = \mathbb{P}(S(\boldsymbol{X}, Y) \leq r \mid \boldsymbol{X}, R = 1)$ is the global CDF of the conformal score,
  - $\overline{\eta}(\boldsymbol{X}) = \mathbb{P}(R = 0 \mid \boldsymbol{X})/\mathbb{P}(R = 1 \mid \boldsymbol{X})$ is the global missingness risk ratio,
  - and $q_0(\boldsymbol{X}) = \mathbb{P}[T = 0 \mid \boldsymbol{X}, R = 0]$ is the target-site propensity.
- However, it will often be unreasonable to assume that the CCOD in practice...

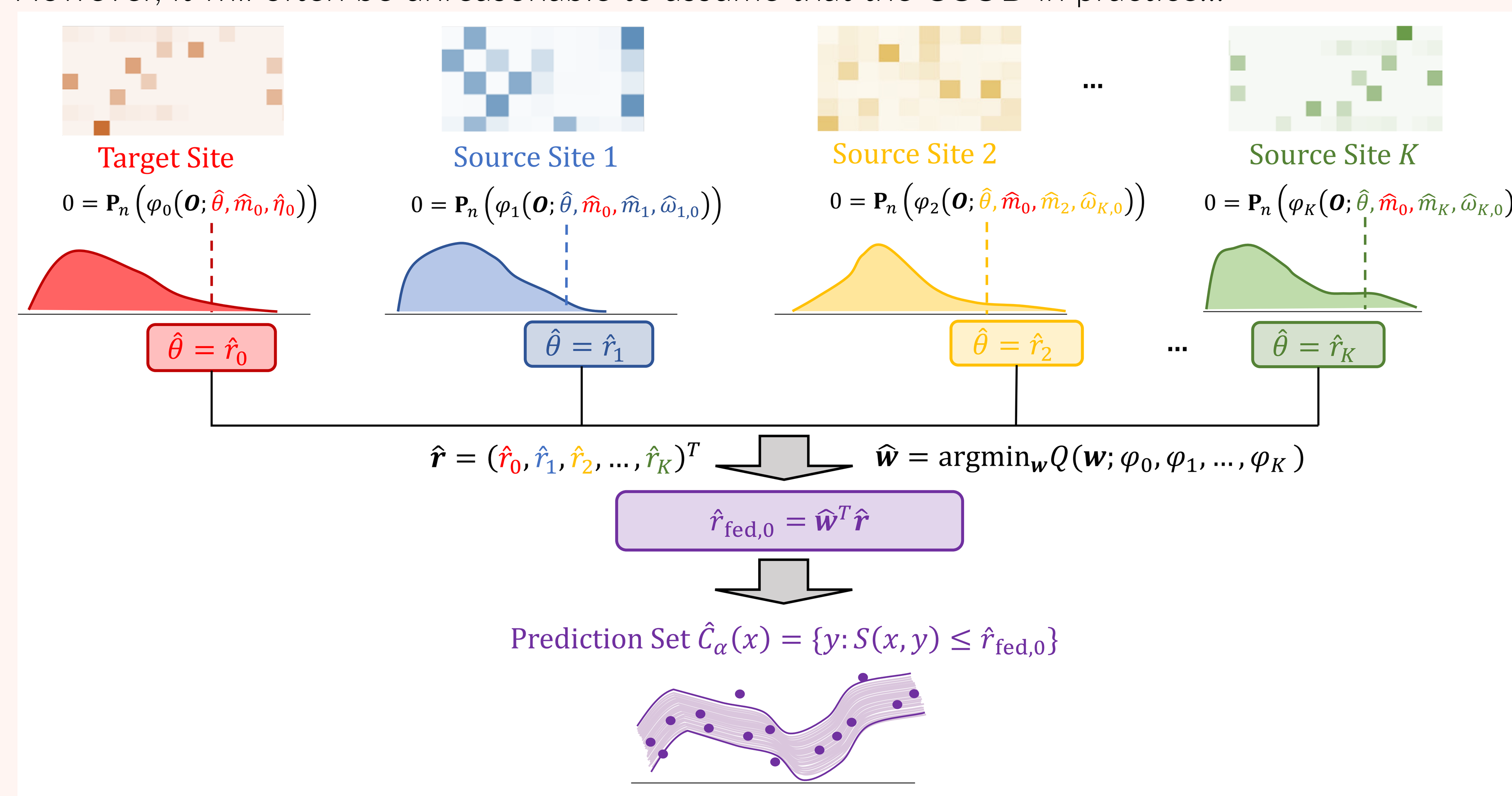


Figure 2. The Proposed Robust Algorithm for Heterogeneous Conditional Outcomes in Multi-Source Data.

For a source site k, the IF of r_0 is given by

$$\frac{I(T = 0, R = 0)}{\mathbb{P}(T = 0, R = 0)}[m_0(r_0, \boldsymbol{X}) - (1 - \alpha)] + \frac{I(T = k, R = 1)}{\mathbb{P}(T = k, R = 1)}\omega_{k,0}(\boldsymbol{X})[I(S(\boldsymbol{X}, Y) \leq r_0) - m_k(r_0, \boldsymbol{X})]$$
$$:= \varphi_k(\mathcal{O}; r_0, m_0, m_k, \omega_{k,0}),$$

where
  - $m_k(r, \boldsymbol{X}) = \mathbb{P}(S(\boldsymbol{X}, Y) \leq r \mid \boldsymbol{X}, T = k, R = 1)$ is the CDF of the conformal score in site k,
  - and $\omega_{k,0}(\boldsymbol{X}) = \dfrac{p(\boldsymbol{X} \mid T = 0, R = 0)}{p(\boldsymbol{X} \mid T = k, R = 1)}$ is a density ratio.
- Limited data sharing: data sharing only comes from the estimation of the density ratio $\omega_{k,0}$. This can be done with the passing of only coarse summary statistics[1].

Data-Adaptive Aggregation

- First compute the discrepancy measures $\widehat{\chi}_k^2 = (\widehat{r}_0 - \widehat{r}_k)^2$.
- Next solve for federated weights $\widehat{\boldsymbol{w}} = (\widehat{w}_0, \widehat{w}_1, \ldots, \widehat{w}_{K-1})$ that minimize the following loss:

$$Q(\boldsymbol{w}) = \mathbb{P}_n\left[\left\{\underbrace{\varphi_0(\mathcal{O}; \widehat{r}_0, \widehat{m}_0, \widehat{\eta}_0)}_{\text{Target IF}} - \sum_{k=1}^{K-1} w_k \underbrace{\varphi_k(\mathcal{O}_i; \widehat{r}_0, \widehat{m}_0, \widehat{m}_k, \widehat{\omega}_{k,0})}_{\text{Source IF}}\right\}^2\right] + \frac{1}{n}\lambda \sum_{k=1}^{K-1} |w_k| \widehat{\chi}_k^2,$$

subject to $0 \leq w_k \leq 1$, for all $k \in \{0, 1, \ldots, K-1\}$, and $\sum_{k=0}^{K-1} w_k = 1$, and λ is a tuning parameter chosen by cross-validation.
- Then compute $\widehat{r}_{0,\text{fed}}$ as the weighted average of the site-specific quantiles: $\widehat{r}_{0,\text{fed}} = \sum_{k=0}^{K-1} \widehat{w}_k \widehat{r}_k$.
- Finally, the federated prediction interval is defined as $\widehat{C}_\alpha^{\text{fed}}(\boldsymbol{X}) = \{y \in \mathbb{R} : S(\boldsymbol{X}, y) \leq \widehat{r}_{0,\text{fed}}\}$.

Numerical Experiments

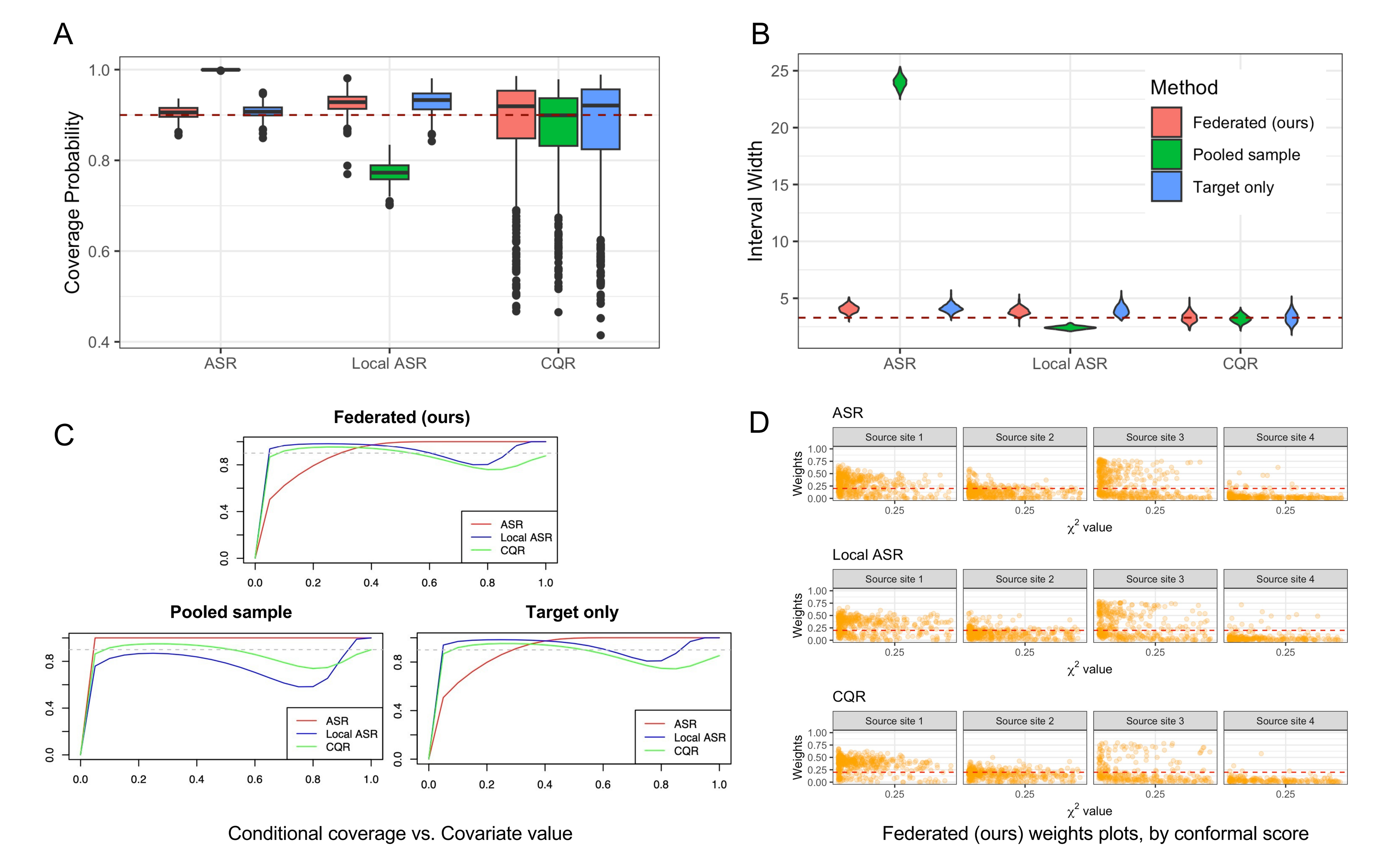


Figure 3. Results by one representative case of in total 162 scenarios of our simulation. We varied: sample sizes n_k, levels of covariate shift, types of outcome errors, levels of concept (outcome) shift, and conformal scores. This case: $K = 5$ sites, $n_k = 1000$ for $k = 0, \ldots, 4$, strongly heterogeneous covariate shift, heteroskedasticity, and strong violation of CCOD.

Concluding Remarks

- We proposed a method to obtain valid prediction intervals for missing outcome data in a target site while exploiting information from multiple potentially heterogeneous sites.
- Marginal coverage properties of conformal prediction methods and builds on modern semiparametric efficiency theory and federated learning for more robust and efficient uncertainty quantification.
- Future research: **Covariate-adaptive ensemble weights** for aggregating information → oracle efficiency. Toward different notions of **conditional coverage**, etc.

Presenting author: Larry Han

Email: lar.han@northeastern.edu