FEDERATED CONFORMAL PREDICTION FOR RELIABLE BRAIN DISORDER SUBTYPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) enables multi-site Functional Magnetic Resonance Imaging (fMRI) analyses to jointly learn shared brain connectivity patterns associated with neurological disorders while adapting to site-specific phenotypic and scanner variations. However, these frameworks often overlook uncertainty quantification for individual model predictions. In neuroimaging, models must alert clinicians when diagnostic predictions are uncertain or have low confidence. This becomes especially important for rare brain disorder subtypes, where limited training data can lead to unreliable yet overconfident predictions masked by high overall model accuracy. To address these issues, we propose BrainFCP (Brain Federated Conformal Prediction), a novel framework that applies Conformal Prediction (CP) in a federated multi-site fMRI setting to provide site-wise and subtype-specific uncertainty estimates with theoretical guarantees. We extensively evaluated our framework on two neurodevelopmental disorders: Attention-Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD), using the ADHD-200 and ABIDE-I resting-state fMRI datasets, respectively. Our results demonstrate BrainFCP's ability to construct valid personalized subtype-level prediction sets for each site in a federated setting.

1 INTRODUCTION

fMRI data are costly to acquire, limited in size, and distributed across multiple sites and institutions (Chan et al., 2022). Deep learning-based fMRI analysis has prompted the use of FL that enables multiple imaging sites to jointly train models while keeping raw data local. This helps reduce privacy concerns through decentralized training. However, data distributions can vary substantially across sites due to factors such as age, gender, and scanner artifacts. Among different FL approaches, personalized FL has therefore been widely adopted, as it can adapt to such site-specific variations while benefiting from collaborative learning (Yang et al., 2023; Zheng et al., 2026).

Existing FL frameworks in neuroimaging often overlook the quantification and communication of uncertainty in individual predictions (Kompa et al., 2021). Uncertainty estimation measures how reliable or confident a deep learning model is in its prediction or classification. As these models may support clinicians in diagnosing and treating neurological disorders, they should be able to abstain from making predictions or alert the user when uncertainty for a patient is high. However, estimating predictive uncertainty in FL is challenging because data across sites are often non-i.i.d. (not independent and identically distributed). Several methods have been proposed to address this issue (Ovadia et al., 2019; Zhang et al., 2022), among which CP is particularly attractive because it is model-agnostic and distribution-free. CP converts a model's point estimate (prediction) into a prediction set. In multi-class classification, this set may contain one or more possible classes. These prediction sets come with a coverage guarantee, meaning that the true outcome is included in the set with a specified probability (Vovk et al., 2005; Tibshirani et al., 2019). Conformal prediction has performed well as an uncertainty estimation technique in FL (Koutsoubis et al., 2024). However, extending Federated Conformal Prediction (FCP) to multi-site fMRI data introduces unique challenges. In our setting, existing FCP methods (Plassier et al., 2023; Min et al., 2026) estimate the similarity between source and target subjects using whole-brain representations. In practice, these representations can mix disease-related signals with scanner-related artifacts (Yamashita et al., 2019). As a result, useful source subjects may be downweighted due to scanner differences, even when they share relevant disease patterns with the target subject. Another key issue is that a CP

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

method may provide a model with prediction sets that achieve, say, 80% coverage when distinguishing controls from disorder patients, yet fail to maintain the same coverage for rarer subtypes. Common subtypes may be over-covered, while atypical subtypes fall well below the target coverage. Such a model may appear reliable overall while remaining unreliable for rare subtypes.

To address these challenges, we introduce BrainFCP (Brain Federated Conformal Prediction), a novel framework that constructs valid prediction sets with both site-wise marginal coverage and subtype-conditional coverage in a multi-site fMRI setting. BrainFCP is extensively evaluated across five FL backbones on two multi-site rs-fMRI datasets covering two neurodevelopmental disorders, ADHD and ASD. We show that BrainFCP consistently achieves near-nominal coverage for rare disorder subtypes and outperforms state-of-the-art FCP methods, supporting its potential clinical utility. Thus, the contributions of this paper are threefold:

- We introduce a site-effect decomposed density ratio estimator that measures source–target similarity using only the biological component of the learned representation. This reduces the influence of scanner differences when weighting biologically relevant source subjects.
- We propose a subtype-aware personalized source mixture that assigns test- and subtype-specific weights to source sites, replacing the sample-size based pooling used in existing methods. We show that the choice of source mixture affects estimation precision but not validity.
- We develop a subtype-stratified calibration scheme with finite-sample subtype-conditional coverage guarantees, together with threshold pooling for rare subtypes with an explicitly bounded coverage loss.

2 Related Work

The earliest work on FCP in medical imaging proposed an aggregation algorithm based on quantile averaging, but without theoretical coverage guarantees (Lu & Kalpathy-Cramer, 2021). This approach was later extended by assuming that the test distribution is a mixture of the training distributions (Lu et al., 2023), with marginal coverage guaranteed under conditional exchangeability across clients. However, this assumption may not hold in personalized FL, where each client can follow a different test distribution. DP-FedCP (Plassier et al., 2023) addresses this heterogeneity by accounting for label shift through weighted quantiles of client non-conformity scores over multiple communication rounds. A later method improves efficiency by computing importance weights using density ratios (Plassier et al., 2024). However, its coverage depends on accurate density ratio estimation, making it sensitive to estimation errors.

To guarantee marginal coverage regardless of the quality of density ratio estimation, PFCP (Min et al., 2026) was proposed. PFCP with localization achieves marginal coverage for a target client under cross-client data heterogeneity by incorporating information from other source distributions to improve local prediction sets without sharing raw data. This is useful when target data are limited and other source distributions contain relevant information about the target. Such a setting is particularly common in neuroimaging, where hospitals may have different numbers of patients across neurological disorder subtypes. Information from other hospitals can therefore help improve estimation for underrepresented subtypes. However, PFCP implicitly assumes alignment between source and target data distributions, which may not hold for multi-site fMRI data. We address these limitations by building on PFCP to model cross-site heterogeneity and improve density ratio estimation for neuroimaging applications. To our knowledge, this is the first work to extend FCP to personalized local settings in neuroimaging.

3 Methodology

Consider BrainFCP framework with K source sites, indexed by $k \in [K]$, and a target site T. The data distributions at source site k and target site T are denoted by P_k and P_T, respectively. Each source site k contains n_k subjects whose data remain local. Each subject is represented as (x, y), where x denotes brain functional connectivity matrix. Formally, x is defined as the Pearson correlation between a pair of brain regions of interest (ROIs). The label $y \in Y = \{1, \ldots, C\}$ denotes the diagnostic label.

BrainFCP is agnostic to the choice of FL backbone h_ψ. Inspired by (Yamashita et al., 2019), we factorize the learned representation into biological and scanner-specific components: $h_\psi(x) = z = (z_b, z_{sc}) \in \mathbb{R}^{d_b} \times \mathbb{R}^{d_s}$. The classification head operates only on the biological representation z_b. We define the non-conformity score as $s(x, y) = 1 - p(y \mid z_b(x)) \in [0, 1]$. A larger score indicates that label y is less consistent with the model's prediction for subject x. Given a miscoverage level $\alpha \in (0, 1)$, the objective is to construct a prediction set $C_\alpha(x) \subseteq Y$ for a test subject at target site T that contains the true subtype with probability at least $1-\alpha$, using information from K collaborating source sites without sharing raw data. The set size reflects predictive uncertainty: a singleton indicates a confident prediction, while multiple labels indicate ambiguity between subtypes. An empty set indicates that the subject is not well supported by the calibration data.

3.1 Site-Effect Decomposition for Density Ratio Estimation

Existing FCP methods determine the contribution of each source subject by comparing its whole-brain representation with that of the target subject. In multi-site fMRI, these representations can mix disease-related signals with scanner-specific effects. As a result, a biologically relevant source subject may be down-weighted simply because it was scanned at a different site. We address this limitation by designing a density ratio estimator that decomposes site-effects.

Definition 1. To quantify how much a source subject should be up- or down-weighted when approximating the target distribution, we define a density ratio. For subtype c, the density ratio is given by:

$$r_b(z_b, c) \;=\; \frac{f_T(z_b \mid c)}{f_{mix}(z_b \mid c)} \tag{1}$$

$f_T(\cdot)$ denotes the density of a biological representation z_b for subtype c at the target site. Similarly, $f_{\text{mix}}(\cdot)$ denotes the density of the same representation in the combined source-site mixture. Instead of estimating the two densities separately, we train a classifier p_θ using only z_b to distinguish target and pooled-source distributions. This is then represented as:-

$$\hat{r_b}(z_b, c) \;=\; \frac{p_\theta(z_b, c)}{1 - p_\theta(z_b, c)} \tag{2}$$

Instead of fitting a separate density ratio estimator for each target-source pair, we train a single conditional estimator that also takes a learned site embedding e_k as input. e_k identifies source site k. This allows one shared network to learn common structure across sites while adapting to site-specific differences. This estimator is defined as $\hat{r_b}\big(z_b, c, e_k\big)$.

Assumption 1. The non-conformity score depends on the input only through the decomposed biological representation: $S \perp z_{sc} \mid (z_b, Y)$. Assumption 1 holds by construction, since the classifier used to compute the non-conformity score takes only z_b as input.

Assumption 2. For every site k and label y, $P_T(y \mid z_b) = P_k(y \mid z_b)$. Assumption 2 holds when the disorder subtype has the same pathophysiological characteristics across all sites. Therefore, sites may differ in the subjects they observe, but not in how the diagnosis relates to the biological representation.

Proposition 1. Under Assumptions 1 and 2, for every non-conformity score $s \in \mathbb{R}$ and every subtype c,

$$F_T(s \mid c) \;=\; \mathbb{E}_{Z \sim f_{\text{mix}}}\Big[\mathbb{1}\{S(Z) \le s\}\; r_b(z_b, c)\Big] \tag{3}$$

where $F_T(\cdot \mid c)$ denotes the target cumulative distribution function of the non-conformity scores. Reweighting source subjects using the density ratio defined in Eq. (1) makes the weighted source score distribution match the target score distribution. Source subjects that are more biologically similar to the target receive greater weight. Proposition 1 supports our claim that scanner-related representations are not required to transport the score distribution. Density ratio can therefore be estimated in the d_b space, making statistical estimation more feasible.

3.2 SUBTYPE-AWARE SOURCE MIXTURE

In standard FCP methods such as PFCP, source sites are pooled in proportion to their sample size. This can be problematic in multi-site fMRI, as a large but poorly matched site may dominate the mixture, while a smaller but more relevant site receives little weight. We replace these fixed weights with test- and subtype-specific weights.

Definition 2. Given z_b and density ratio estimator $\hat{r_b}(\cdot)$ defined in Eq. (2), the personalized mixture weights are defined as:

$$\pi_{k,c}(x) \;=\; G_\phi\big(z_b(x),\; p_k^c,\; \hat{r_b}(z_b, c, e_k),\; u_k^c\big), \qquad \pi_{k,c}(x) \geq 0, \quad \textstyle\sum_k \pi_{k,c}(x) = 1 \tag{4}$$

where p_k^c denotes the subtype-c prototype at site k. It represents the mean biological representation of subtype-c at that site. The term u_k^c denotes the uncertainty in the estimated density ratio for site k and subtype c. G_ϕ is a learnable network that takes all these terms as input and outputs the corresponding source weight.

We compute u_k^c as the posterior variance of the estimated log-density ratio: $u_k^c = \mathrm{Var}\left[\log \hat{r}_b(\cdot)\right]$. If G_ϕ assigns a large weight to a source with a highly uncertain density ratio estimate, we use u_k^c to shrink that weight toward a safer uniform weighting. We then shrink these weights:-

$$\tilde{\pi}_{k,c}(x) \;=\; (1-\beta_k)\,\pi_{k,c}(x) \;+\; \beta_k \cdot \frac{1}{K}, \qquad \beta_k \;=\; \frac{u_k^c}{u_k^c + \tau} \tag{5}$$

where τ is a hyperparameter controlling shrinkage. Lastly, following Definition 2, the modified mixture and density ratio are given as:-

$$f_{\mathrm{mix}}^{c,x} \;=\; \sum_{k=1}^{K} \pi_{k,c}(x)\, f_k, \qquad \tilde{r_b}(z_b, s, c) \;=\; \frac{f_T(z_b, s \mid c)}{f_{\mathrm{mix}}^{c,x}(z_b, s \mid c)} \tag{6}$$

It is worth noting that Eq.(6) differs from Eq.(1) by incorporating non-conformity score s. This is important because CP ultimately depends on the distribution of these scores, not only on z_b. Eq.(6) therefore measures how common a given biological representation and non-conformity score are in the target distribution relative to the personalized source mixture.

Theorem 1. Let $\hat{F}(s \mid x)$ denote the M-sample importance-sampling estimator of the target conditional score CDF, obtained from Eq.(3) using $f_{\mathrm{mix}}^{c,x}$. Then, provided $f_{\mathrm{mix}}^{c,x}$ adequately covers $f_T(\cdot \mid c)$:-

1. $\mathbb{E}[\hat{F}(s \mid x)] = F_T(s \mid c)$ for any valid choice of mixture weights.
2. Its variance is bounded by $\mathrm{Var}\big[\hat{F}(s \mid x)\big] \leq \frac{1}{M}\left(1 + \chi^2\big(f_T \,\|\, f_{\mathrm{mix}}^{c,x}\big)\right)$

Here, $\chi^2(\cdot)$ measures the mismatch between the target and source mixture distributions. The key interpretation of Theorem 1 is that different source-site weights $\pi_{k,c}(x)$ yield unbiased target score CDF estimates but with different variance, provided the source mixture covers the target distribution. Thus, the source mixture affects estimation efficiency only.

3.3 SUBTYPE-STRATIFIED CALIBRATION

A single calibration threshold guarantees coverage only on average across the target population. This can hide disorder subtype-level failures, where a majority subtype is over-covered while a rare subtype falls below the target coverage. To address this, we calibrate separately for each subtype using a label-conditional construction (Mondrian cells).

Let $\mathcal{D}_c^T = \{i \in \mathcal{D}_{\mathrm{cal}}^T : Y_i = c\}$ denote the target calibration subjects belonging to subtype-c and let $n_c = |\mathcal{D}_c^T|$ denote their size. For each subtype c, the threshold is given as:-

$$q_c \;=\; \text{the } \Big\lceil (1-\alpha)(n_c+1) \Big\rceil \text{-th smallest value of } \big\{\, V_i \;:\; i \in \mathcal{D}_c^T \,\big\} \tag{7}$$

where V_i represents the position of subject i's conformity score within the estimated target-like score distribution. For a new subject x', the prediction set is computed as:-

$$C_\alpha(x') \;=\; \{\, y \in Y \;:\; V^y(x') \;\le\; q_y \,\} \tag{8}$$

Each candidate subtype is evaluated against its own threshold rather than a shared threshold. This allows prediction sets to be constructed independently for each subtype.

Theorem 2. Suppose that target calibration set $\mathcal{D}^T_{\text{cal}}$ is disjoint from all data used to train the encoder h_ψ, density ratio estimator $\hat{r_b}$ and mixture network G_ϕ. Let the target calibration subjects $(X_i, Y_i)_{i\in\mathcal{D}^T_{\text{cal}}}$ and test subject (X_{n+1}, Y_{n+1}) be i.i.d. samples from P_T. Then, for each subtype-c:-

$$\mathbb{P}\Big(Y_{n+1} \in C_\alpha(X_{n+1}) \;\Big|\; Y_{n+1} = c \Big) \;\ge\; 1-\alpha \tag{9}$$

If no two transformed conformity scores V_i have exactly the same value, the coverage is upper bounded by $P(Y \in C \mid Y = c) < 1 - \alpha + (n_c + 1)^{-1}$.

Theorem 2 establishes the finite-sample subtype-conditional validity of BrainFCP. The coverage guarantee comes from calibrating only on subjects from the target site and if target calibration subjects and the new target test subject are coming from the same distribution. It does not depend on the quality of the site-effect decomposition, density ratio estimation, or source-target matching. Poor source mixture weighting cannot invalidate the guarantee. It can only result in less efficient and wider prediction sets. Sections 3.1–3.2 therefore does not affect validity. They should instead be viewed as improving the efficiency and sharpness of the prediction sets.

Proposition 2. Subtype-wise calibration protects rare subtypes by enforcing separate coverage guarantees. However, for a rare subtype c at a small target site, the calibration size n_c may be small, resulting in a noisier or potentially undefined subtype-specific threshold. We therefore shrink each subtype-specific threshold toward the pooled threshold q_p using all target calibration subjects. Specifically:

$$q_c^{\text{sh}} \;=\; \lambda_c\, q_c \;+\; (1-\lambda_c)\, q_p, \quad \lambda_c \;=\; \frac{n_c}{n_c + \tau_0} \tag{10}$$

However, threshold pooling comes at the cost of reduced subtype-specific validity. Let F_c and F_p denote the score distributions for subtype c and the pooled target population, respectively. We define $\Delta_c = \text{TV}(F_c, F_p)$, which measures the difference between the subtype-specific score distribution and the overall target distribution. Then, the prediction set constructed using q_c^{sh} is bounded by:-

$$\mathbb{P}\big(Y \in C_\alpha(X) \mid Y = c\big) \;\ge\; 1 - \alpha - (1-\lambda_c)\,\Delta_c \tag{11}$$

Remark 1. Theorem 2 applies to any measurable clinical grouping defined from the subject, with coverage guaranteed conditionally on that grouping. Here, the grouping is the disorder subtype. Finer groups, such as dynamic Functional Connectivity (FC) phenotypes or age categories, are also possible but result in calibration sets of different sizes. Thus, more specific groups provide more localized guarantees, with the available calibration data varying across groups.

Remark 2. The prediction set constructed from Eq.(8) may be empty when no candidate subtype satisfies its threshold. An empty set indicates that the subject is atypical for all calibrated subtypes and may lie outside the calibration distribution. We therefore treat empty sets as explicit out-of-distribution flag, which is a safer outcome in clinical settings.

Remark 3. All coverage guarantees require subject-level exchangeability only within the target site. We further note that dynamic FC representations should not be treated as exchangeable at the sliding-window level, since windows from the same subject are temporally correlated. Static FC naturally produces one FC matrix per participant.

Remark 4. Source sites may differ in the subjects they recruit and in their scanning protocols, leading to covariate shift. However, the relationship between the biological representation and subtype label is assumed to remain consistent across sites. If a subtype presents differently at the target site, weights estimated from unlabeled scans cannot capture this difference.

4 EXPERIMENTS

4.1 DATASETS AND PREPROCESSING

We evaluated BrainFCP on two public multi-site rs-fMRI datasets, demonstrating its coverage validity and generalizability.

ADHD-200 (Brown et al., 2012): From the ADHD-200 dataset, we used rs-fMRI scans from 285 subjects with ADHD and 341 healthy controls across five sites. We selected NYU (99 NC / 123 ADHD), OHSU (42 NC / 37 ADHD), PKU (116 NC / 78 ADHD), NI (23 NC / 25 ADHD), and KKI (61 NC / 22 ADHD). WUSTL (61 NC / 0 ADHD) was excluded because it contained only controls and no ADHD subjects. There are 3 subtypes of ADHD: Hyperactive-Impulsive, Inattentive and Combined. Detailed site-wise breakdown of ADHD subtypes is provided in Figure 1 (a). We obtained the preprocessed ADHD-200 data using the Athena pipeline (Bellec et al., 2017).

ABIDE-I (Di Martino et al., 2014): The Autism Brain Imaging Data Exchange I (ABIDE-I) dataset contains rs-fMRI scans from individuals diagnosed with ASD and typically developing controls, collected from 17 sites. We selected sites: NYU (105 NC / 79 ASD), KKI (27 NC / 28 ASD), OLIN (16 NC / 19 ASD), SDSU (24 NC / 24 ASD), UCLA (41 NC / 43 ASD), UM (74 NC / 66 ASD) and YALE (28 NC / 28 ASD). In ABIDE-I, ASD subtypes are defined according to the DSM-IV-TR criteria (Association et al., 2000), which classify ASD into three categories: Autistic Disorder, Asperger's syndrome (AS), and Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS). Remaining sites were excluded because they did not contain subjects from all three ASD subtypes. Detailed site-wise breakdown of ASD subtypes is provided in Figure 1 (b). ABIDE-I dataset was preprocessed using the C-PAC pipeline from the Preprocessed Connectomes Project.

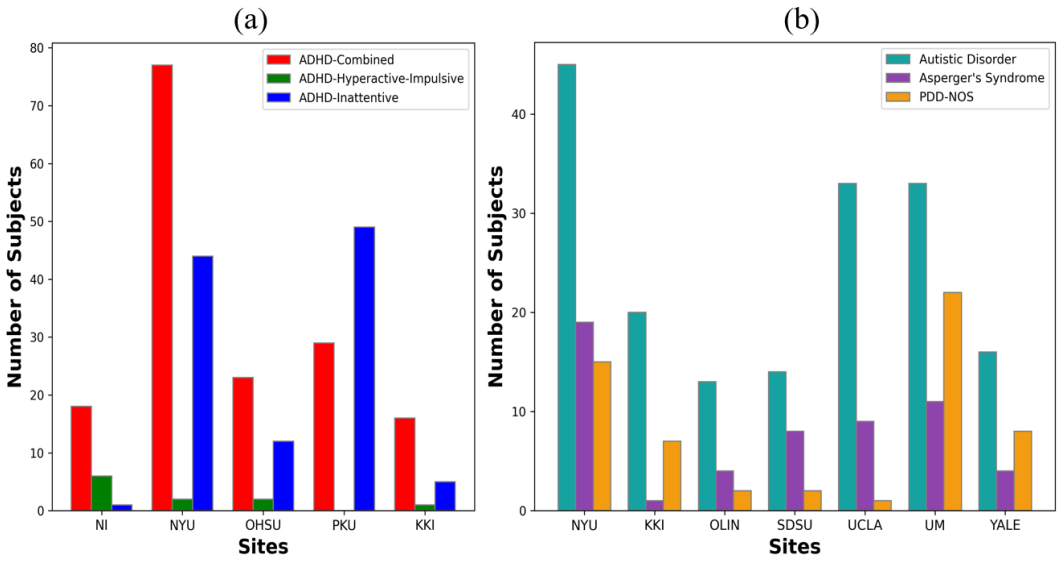

Figure 1: Frequency distribution of subjects by subtype across selected sites in (a) ADHD-200 dataset and (b) ABIDE-I dataset.

4.2 BASELINES

We conducted experiments using several state-of-the-art FL baselines, including both generic and fMRI-specific methods: FedAvg (McMahan et al., 2017), FedProto (Tan et al., 2022), FedBrain (Yang et al., 2023), FedGST (Mao et al., 2023) and GAFD (Wang et al., 2025). FedAvg performs weighted parameter averaging across clients to construct a global model, while FedProto exchanges class-level feature prototypes instead of model parameters. FedBrain (Yang et al., 2023) uses a GNN-based FL framework that incorporates brain connectome properties for disease prediction from multimodal imaging data. FedGST uses multiple sliding-window FC matrices internally, but we modify it to produce one aggregated FC matrix per subject. GAFD applies graph augmentation

Table 1: Comparison of marginal coverage and average prediction set size on the ADHD-200 dataset. Results are reported as mean ± standard deviation.

	Models	PFCP	FedCP	DP-FedCP	CP-DH	BrainFCP
Marginal Coverage	FedAvg	0.84 ± 0.02	0.87 ± 0.03	0.86 ± 0.03	0.84 ± 0.03	$\mathbf{0.82 \pm 0.03}$
	FedProto	0.84 ± 0.03	0.86 ± 0.04	0.87 ± 0.02	0.85 ± 0.04	$\mathbf{0.82 \pm 0.04}$
	FedBrain	0.85 ± 0.02	0.85 ± 0.03	0.85 ± 0.03	0.84 ± 0.02	$\mathbf{0.82 \pm 0.02}$
	FedGST	0.84 ± 0.02	0.86 ± 0.03	0.85 ± 0.04	0.83 ± 0.03	$\mathbf{0.81 \pm 0.02}$
	GAFD	0.83 ± 0.01	0.85 ± 0.02	0.84 ± 0.03	0.83 ± 0.03	$\mathbf{0.80 \pm 0.02}$
Average Setsize	FedAvg	2.12 ± 0.06	2.79 ± 0.08	2.63 ± 0.09	2.38 ± 0.08	$\mathbf{2.03 \pm 0.05}$
	FedProto	2.09 ± 0.06	2.65 ± 0.07	2.54 ± 0.08	2.21 ± 0.07	$\mathbf{1.83 \pm 0.07}$
	FedBrain	1.98 ± 0.07	2.58 ± 0.08	2.48 ± 0.08	2.17 ± 0.07	$\mathbf{1.66 \pm 0.07}$
	FedGST	1.88 ± 0.08	2.51 ± 0.07	2.39 ± 0.07	2.05 ± 0.07	$\mathbf{1.57 \pm 0.06}$
	GAFD	1.74 ± 0.06	2.40 ± 0.07	2.33 ± 0.07	1.98 ± 0.05	$\mathbf{1.52 \pm 0.06}$

Table 2: Comparison of marginal coverage and average prediction set size on the ABIDE-I dataset. Results are reported as mean ± standard deviation.

	Models	PFCP	FedCP	DP-FedCP	CP-DH	BrainFCP
Marginal Coverage	FedAvg	0.84 ± 0.06	0.88 ± 0.05	0.85 ± 0.04	0.84 ± 0.06	$\mathbf{0.81 \pm 0.05}$
	FedProto	0.83 ± 0.05	0.87 ± 0.05	0.85 ± 0.05	0.84 ± 0.05	$\mathbf{0.81 \pm 0.03}$
	FedBrain	0.84 ± 0.05	0.86 ± 0.04	0.84 ± 0.04	0.83 ± 0.04	$\mathbf{0.80 \pm 0.03}$
	FedGST	0.83 ± 0.04	0.86 ± 0.04	0.84 ± 0.05	0.84 ± 0.03	$\mathbf{0.80 \pm 0.04}$
	GAFD	0.82 ± 0.03	0.85 ± 0.05	0.83 ± 0.03	0.83 ± 0.03	$\mathbf{0.79 \pm 0.03}$
Average Setsize	FedAvg	2.22 ± 0.07	2.81 ± 0.06	2.76 ± 0.07	2.44 ± 0.07	$\mathbf{1.92 \pm 0.06}$
	FedProto	2.14 ± 0.07	2.71 ± 0.06	2.66 ± 0.06	2.38 ± 0.05	$\mathbf{1.89 \pm 0.07}$
	FedBrain	2.05 ± 0.06	2.62 ± 0.05	2.57 ± 0.05	2.27 ± 0.06	$\mathbf{1.81 \pm 0.07}$
	FedGST	2.03 ± 0.05	2.56 ± 0.05	2.36 ± 0.05	2.10 ± 0.05	$\mathbf{1.67 \pm 0.05}$
	GAFD	1.90 ± 0.06	2.39 ± 0.05	2.24 ± 0.05	2.01 ± 0.06	$\mathbf{1.61 \pm 0.06}$

and knowledge distillation to static FC networks for multi-site aggregation. To evaluate BrainFCP, we compared it against existing FCP methods, including PFCP (Min et al., 2026), FedCP (Lu et al., 2023), DP-FedCP (Plassier et al., 2023), and CP-DH (Plassier et al., 2024).

4.3 Implementation Details

Experiments followed a leave-one-site-as-target protocol, with each site serving as the target in turn. At each target site, subjects were split into training, calibration, and test sets. Subjects were split using stratified random sampling by diagnostic subtype. For rare subtypes, covariate-constrained randomization was used to ensure balanced representation in the training, calibration, and test sets wherever possible. The training set was used to estimate the proposed density ratio and source mixture weights. Source sites were used to train the FL models, after which the model weights were frozen. FC matrices were constructed using Pearson correlations between the mean time series of each pair of ROIs. Power atlas (Power et al., 2011) was used to define 264 ROIs. All FL models were implemented in Python 3.10 and PyTorch 2.1 and trained on an NVIDIA A100 GPU. Unless otherwise specified, default hyperparameters were used for all baseline FL models. The miscoverage level was set to $\alpha = 0.2$ for all experiments. The label set Y includes controls and three disorder subtypes corresponding to ADHD or ASD. We compare BrainFCP with other FCP methods using the prediction set C_α based on three criteria: marginal coverage, subtype-conditional coverage and average prediction setsize. For each target site, results were first averaged across five seeds. We then report the mean ± standard deviation of these site-wise results across all target sites.

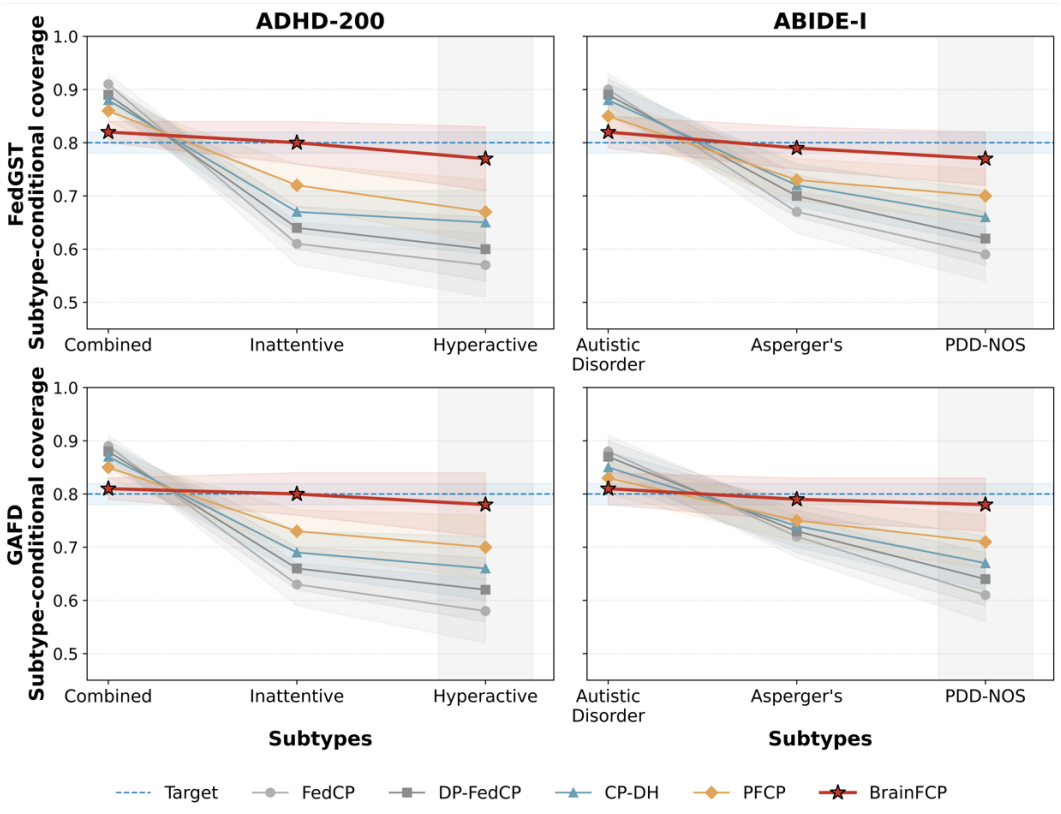

Figure 2: Subtype-conditional coverage of FCP methods using FedGST (top) and GAFD (bottom) backbones on ADHD-200 and ABIDE-I dataset. The dashed line indicates the target coverage ($1 - \alpha = 0.80$) and shaded bands show ± 1 standard deviation. Grey region highlights the rarest subtype.

4.4 Performance Comparison with Baselines

Tables 1 and 2 summarize the marginal coverage and average prediction set size results on the ADHD-200 and ABIDE-I datasets, respectively. Baseline FCP methods consistently overcover the nominal marginal coverage of $1 - \alpha = 0.80$, resulting in larger prediction sets without improving the coverage guarantee. In contrast, BrainFCP achieves coverage closest to the target while producing the smallest prediction sets across all five FL backbones. Compared to PFCP, the strongest baseline, BrainFCP reduces the average set size by 4–17% on ADHD-200 and 12–18% on ABIDE-I. These improvements are consistent across both generic (FedAvg, FedProto) and fMRI-specific (FedBrain, FedGST, GAFD) backbones, demonstrating that BrainFCP is backbone-agnostic. Across all FCP methods, stronger backbones generally produce smaller prediction sets, with FedGST and GAFD yielding the smallest sets, likely due to more informative FC representations and sharper non-conformity scores.

To evaluate subtype-conditional coverage on both datasets, we selected the two best-performing FL backbones, FedGST and GAFD. Based on Eq.(7), we compute a subtype-specific threshold q_c for each subtype. The results are shown in Figure 2. Figure 2 shows that marginal coverage can conceal substantial subtype-level coverage failures. BrainFCP substantially improves rare-subtype coverage compared with existing FCP methods. Baseline methods tend to over-cover the majority subtype while under-covering rarer subtypes. On ADHD-200, for example, FedCP achieves a marginal coverage of approximately 0.85 with both FedGST and GAFD, yet covers fewer than 60% of subjects in the Hyperactive subtype. PFCP, which also leverages cross-site information, achieves only about 0.67 coverage for the rarest subtype. In contrast, BrainFCP maintains coverage close to the target for well-represented subtypes, with only a small reduction for the rarest subtype. This is expected because subtypes such as ADHD Hyperactive and ASD PDD-NOS contain relatively few subjects.

Table 3: Comparison of matched coverage across all FCP methods using GAFD as the FL backbone on the ADHD-200 and ABIDE-I datasets. Minimum SCC denotes the minimum subtype-conditional coverage. Results are reported as mean ± standard deviation across target sites.

Dataset	Method	α^*	Coverage	Set size	Minimum SCC
ADHD-200	FedCP	0.26	0.81 ± 0.04	1.98 ± 0.07	0.57 ± 0.07
	DP-FedCP	0.25	0.80 ± 0.03	1.90 ± 0.07	0.59 ± 0.06
	CP-DH	0.24	0.79 ± 0.03	1.83 ± 0.06	0.64 ± 0.06
	PFCP	0.22	0.80 ± 0.03	1.71 ± 0.07	0.68 ± 0.06
	BrainFCP	**0.20**	0.80 ± 0.02	$\mathbf{1.52 \pm 0.06}$	$\mathbf{0.78 \pm 0.05}$
ABIDE-I	FedCP	0.27	0.79 ± 0.05	1.96 ± 0.06	0.56 ± 0.07
	DP-FedCP	0.26	0.81 ± 0.03	1.93 ± 0.06	0.60 ± 0.06
	CP-DH	0.26	0.80 ± 0.04	1.85 ± 0.05	0.64 ± 0.06
	PFCP	0.23	0.80 ± 0.03	1.77 ± 0.06	0.69 ± 0.05
	BrainFCP	**0.20**	0.79 ± 0.03	$\mathbf{1.61 \pm 0.06}$	$\mathbf{0.78 \pm 0.05}$

Their calibration therefore relies on the shrunk threshold q_c^{sh}, with the resulting coverage remaining within the bound established in Proposition 2.

4.5 Matched Coverage Analysis

To verify that BrainFCP's smaller prediction sets are not simply due to less conservative coverage, we tune each baseline's miscoverage level α^* to achieve approximately 0.80 marginal coverage. We use the target-site training split for tuning, which is otherwise unused by the baseline FCP methods. We then compare all methods under this matched-coverage setting. The results are summarized in Table 3. To match the target marginal coverage, all baselines required a larger miscoverage level α^*, resulting in lower calibration thresholds. This supports our initial observation from Tables 1 and 2 that the baseline methods over-cover at the nominal level. Under matched coverage, BrainFCP still produces the smallest prediction sets, reducing the average set size by 11.1% and 9.0% relative to PFCP on ADHD-200 and ABIDE-I, respectively. However, matching marginal coverage further reduces the baselines' minimum-subtype coverage. This occurs because loosening a shared threshold disproportionately removes labels for rare subtypes, where the model is less confident. In contrast, BrainFCP maintains a minimum subtype-conditional coverage of 0.78 on both datasets. BrainFCP achieves strong coverage across both datasets, suggesting good generalizability. The ablation results in Table 4 (See Appendix) further support the contribution of its novel components.

5 Discussion and Conclusion

This study presents several key findings. First, the site-effect decomposed density ratio allows BrainFCP to borrow information from biologically similar source subjects while reducing the influence of scanner differences. We further show that the biological representation is sufficient to transport the target score distribution. Second, the subtype-aware source mixture replaces sample-size-based pooling with test- and subtype-specific weights. From Theorem 1, a source mixture closer to the target distribution reduces the variance of the estimated score distribution, leading to tighter thresholds and smaller prediction sets. This helps explain why BrainFCP remains efficient under matched coverage. Finally, baseline methods use a shared threshold that can over-cover majority subtypes while under-covering rarer ones. In contrast, BrainFCP calibrates each subtype separately and maintains stronger subtype-conditional coverage across both datasets.

A key design property of BrainFCP is the separation of validity and efficiency. Coverage is guaranteed through target-site calibration (Theorem 2), while the components introduced in Sections 3.1 and 3.2 affect only efficiency. Thus, imperfect source weighting may widen prediction sets but does not invalidate the coverage guarantee. One limitation of this study is that the rarest subtypes contain too few subjects to estimate reliable subtype-specific thresholds and therefore require threshold pooling. In addition, the ASD subtypes in the ABIDE-I dataset are defined using the DSM-IV-TR taxonomy, which has since been merged into a single autism spectrum in DSM-5 (Association, 2022). Future work will extend BrainFCP to dynamic FC and finer clinical groupings.

REFERENCES

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association Publishing, 2022.

American Psychiatric Association et al. Diagnostic and statistical manual of mental disorders 4th edition, text revision. *Washington, DC: American psychiatric association*, pp. 121–134, 2000.

Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *Neuroimage*, 144:275–286, 2017.

Matthew RG Brown, Gagan S Sidhu, Russell Greiner, Nasimeh Asgarian, Meysam Bastani, Peter H Silverstone, Andrew J Greenshaw, and Serdar M Dursun. Adhd-200 global competition: diagnosing adhd using personal characteristic data can outperform resting state fmri measurements. *Frontiers in systems neuroscience*, 6:69, 2012.

Yi Hao Chan, Wei Chee Yew, and Jagath C Rajapakse. Semi-supervised learning with data harmonisation for biomarker discovery from resting state fmri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 441–451. Springer, 2022.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

Nikolas Koutsoubis, Yasin Yilmaz, Ravi P Ramachandran, Matthew Schabath, and Ghulam Rasool. Privacy preserving federated learning in medical imaging with uncertainty estimation. *arXiv preprint arXiv:2406.12815*, 2024.

Charles Lu and Jayasheree Kalpathy-Cramer. Distribution-free federated learning with conformal predictions. *arXiv preprint arXiv:2110.07661*, 2021.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pp. 22942–22964. PMLR, 2023.

Junbin Mao, Hanhe Lin, Xu Tian, Yi Pan, and Jin Liu. Fedgst: Federated graph spatio-temporal framework for brain functional disease prediction. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1356–1361. IEEE, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Yinjie Min, Chuchen Zhang, Liuhua Peng, and Changliang Zou. Personalized federated conformal prediction with localization. *Advances in Neural Information Processing Systems*, 38:101548–101585, 2026.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pp. 27907–27947. PMLR, 2023.

Vincent Plassier, Nikita Kotelevskii, Aleksandr Rubashevskii, Fedor Noskov, Maksim Velikanov, Alexander Fishkov, Samuel Horvath, Martin Takac, Eric Moulines, and Maxim Panov. Efficient conformal prediction under data heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 4879–4887. PMLR, 2024.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8432–8440, 2022.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Qianqian Wang, Junhao Zhang, Long Li, Lishan Qiao, Pew-Thian Yap, and Mingxia Liu. Graph augmentation guided federated knowledge distillation for multisite functional mri analysis. *Pattern Recognition*, 164:111559, 2025.

Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, Naohiro Okada, et al. Harmonization of resting-state functional mri data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS biology*, 17(4):e3000042, 2019.

Yi Yang, Han Xie, Hejie Cui, and Carl Yang. Fedbrain: Federated training of graph neural networks for connectome-based brain imaging analysis. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pp. 214–225. World Scientific, 2023.

Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized federated learning via variational bayesian inference. In *International Conference on Machine Learning*, pp. 26293–26310. PMLR, 2022.

Yuxuan Zheng, Zifeng Jia, Zihao Guan, Yuzhong Chen, Renjie Zhou, Keith M Kendrick, and Xi Jiang. Toppfed: Topological prototype-enhanced personalized federated learning for neuropsychiatric disorders identification. *IEEE Transactions on Medical Imaging*, 2026.

A Appendix

A.1 Ablation Study

Table 4: Ablation study of BrainFCP components on ADHD-200 using the GAFD backbone ($\alpha = 0.20$). Minimum SCC denotes the minimum subtype-conditional coverage. Results are reported as mean $\pm$ standard deviation across target sites.

Variant	Coverage	Set size	Minimum SCC
BrainFCP (full)	$\mathbf{0.80 \pm 0.02}$	$\mathbf{1.52 \pm 0.06}$	$\mathbf{0.78 \pm 0.05}$
w/o site-effect decomposition	0.80 ± 0.03	1.65 ± 0.06	0.76 ± 0.05
w/o subtype-aware mixture	0.80 ± 0.02	1.59 ± 0.07	0.77 ± 0.06
w/o subtype stratification	0.81 ± 0.02	1.55 ± 0.06	0.66 ± 0.07
w/o decomposition & mixture	0.81 ± 0.02	1.76 ± 0.07	0.76 ± 0.06

To assess the contribution of each component, we perform an ablation study on ADHD-200 using GAFD as the FL backbone. The results are shown in Table 4. All BrainFCP variants maintain marginal coverage close to the target. This confirms that validity comes from target-site calibration, as established in Theorem 2. Removing the site-effect decomposition increases the average prediction set size from 1.52 to 1.65, showing that density ratio estimation in the biological representation space improves efficiency. Removing both the subtype-aware mixture and site-effect decomposition leads to the largest increase in set size, suggesting that the two components are complementary. In contrast, removing subtype stratification has little effect on set size but reduces the minimum subtype-conditional coverage from 0.78 to 0.66. This demonstrates that site-effect decomposition and subtype-aware mixture weighting primarily improve estimation efficiency, while subtype stratification is responsible for reliable coverage of rare subtypes.

A.2 AI-Use Disclosure

In this work, we used generative AI tools only to polish the manuscript and assist with code debugging. All AI-assisted text was reviewed by the authors for accuracy and consistency with the reported results and all AI-assisted code changes were verified and tested. All reported results were produced from our own experiments. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.