PRIVACY CONSIDERATIONS IN MULTI-INSTITUTIONAL DATA HARMONIZATION THROUGH A CATEGORY-THEORETIC LENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Institutions pool data to overcome local scarcity, and filtering-based matching harmonizes data pooling by admitting only samples that align with a reference site. We show it is unsafe for privacy, and that the two failures have a common cause. For an institution/site whose population not harmonized with respect to other institutions, the samples that align best with the reference institution/site are precisely the atypical ones in its own distribution, so the attempt to harmonize data forces each site to contribute an unrepresentative subset of itself. What leaks is then not a patient record but the distributional identity of the institution. We therefore propose pooling specific a ϵ–δ DP metric (DP_{pool}) for data pooling, a task for which privacy has not been studied and is fundamentally different from existing DP setups. Using category theory helps us demonstrate and understand the natural tension between data harmonization and DP_{pool}, proving a structural incompatibility where improving data harmonization (utility) strictly degrades DP_{pool} (privacy) for confounded institutions under data distribution shifts. To resolve this, we propose privacy-budgeted minimax matching (PBMM) that jointly optimizes for distributional alignment, sample efficiency, and differential privacy. Across synthetic, multi-site fMRI (homogeneous institutional distribution) and six-modality medical imaging benchmarks (heterogeneous institutional distribution), PBMM preserves the harmonization of filtering while cutting peak in-domain exposure by over $3\times$ and recovering samples that filtering discards.

1 INTRODUCTION

Modern AI models typically require large-scale datasets to achieve reliable convergence and strong generalization (Yan et al., 2025; Muennighoff et al., 2023). In many real-world settings, particularly in medical imaging, data scarcity remains a fundamental bottleneck, driven by: (i) the high cost of expert curation and annotation (Lawley et al., 2024) and (ii) stringent privacy and governance requirements that limit data sharing across institutions (Kaye, 2012). A natural and widely adopted solution is to pool data across multiple hospitals, cohorts, or sources, which can alleviate limited sample sizes for underrepresented populations or data-scarce domains while reducing the logistical and financial burden of building large datasets.

The feasibility of such cross-institutional pooling is supported both by long-standing statistical methodology, including covariate matching and meta-analysis Mahajan et al. (2021); Kim & Steiner (2016); Lokhande et al. (2022), and by empirical successes in multi-site studies. Yet, aggregating heterogeneous data is not innocuous: in multi-domain settings, simply increasing dataset size can paradoxically degrade overall performance (Gulrajani & Lopez-Paz, 2020; Zhou et al., 2022), introduce fairness uncertainties Mehta et al. (2024), and reduce accuracy on the most challenging or minoritized subgroups (Shen et al., 2024; Roy et al., 2025). These issues are not confined to biomedical applications (Roy et al., 2026; 2025), but also arise in broader machine learning and computer vision pipelines where large training sets are assembled by pooling datasets from diverse sources (Mattioli, 2017). A complementary perspective is offered by scaling theory: Yan et al. (2025) shows that when data is scarce, repeated epochs over a fixed dataset provide diminishing returns at a rate that grows with dataset size. *We need principled pooling rather than blind data accumulation.*

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.

Multi-site analyses conventionally remove nuisance variation through invariant representation learning, but latent-space adjustment cannot address every covariate (Moyer et al., 2020; Luo & Ren, 2024), which motivates intervening in the data itself. Recent filtering-based pooling does this, establishing that what matters is *which* samples are selected, not how many (Shen et al., 2024; Roy et al., 2026). But selection is not a neutral act: a site's contributed subset is exposed at the moment it is admitted, and how unrepresentative that subset is of the site already determines what the pool discloses about it. So pooling must be budgeted for privacy at the point of selection, not audited for it afterwards. Existing notions cannot supply that budget: privacy-enhancing technologies protect the *computation* and leave selection untouched (McMahan et al., 2017; Hatamizadeh et al., 2022), while differential privacy is calibrated to a worst case over hypothetical datasets and enforced by perturbing a training set already chosen, giving no lever on the choice itself. A privacy notion for pooling must therefore be a function of the selection, so that tightening it changes which samples are admitted.

Example 1. *Consider a stylized collaborative chest X-ray study with three hospitals: a large academic center (H_1, $n_1 = 5{,}000$), a regional hospital (H_2, $n_2 = 800$), and a specialist clinic (H_3, $n_3 = 300$). Pooling (H_1, H_2, H_3) operate under a shared (ϵ, δ) privacy budget, ensuring that inclusion of any individual patient cannot be reliably inferred (sequential addition of (H_i) without information leakage). Each radiograph costs* \$8–25 *to annotate, while a privacy violation incurs a* \$100 *penalty per patient, under a total budget of* \$70,000 *set by the sponsors of the study. A naive* include-all *strategy fails due to inclusion cost of all the samples and the privacy cost for information leakage. Filtering techniques filters confounded samples for better performance (utility) while taking heavy penalty for privacy violation. Conversely, a strategy that filters samples aggressively to preserve privacy meets the budget but discards informative samples, harming coverage of key subgroups. Thus, neither alignment-only pooling nor privacy-only filtering is sufficient.*

Therefore in collaborative multi-institutional studies, each site operates under a joint budget that couples annotation cost with our pooling specific privacy constraints (DP_{pool}). This is especially critical in medical imaging, where sensitive attributes (e.g., identity, demographics) must remain protected Kaye (2012). Consequently, including a sample in a pooled dataset consumes both labeling resources and privacy budget, with potential regulatory and financial risks if exceeded. Existing methods largely overlook this three-way trade-off among (i) aligning pooled data to a target distribution, (ii) maximizing statistical utility, and (iii) respecting privacy constraints (see Example 2). The challenge is amplified in data-scarce settings with weak supervision Kazeminia et al. (2023), where sample selection directly impacts causal validity. Prior work Akbar et al. (2025) interprets outcome-invariant augmentation as interventions on the treatment mechanism, which motivates principled, invariant selection strategies across heterogeneous domains (domains with highly dissimilar data distributions). However, aggressive augmentations (e.g., noise or texture perturbations) can distort the data manifold, degrading calibration Müller et al. (2019) and inducing covariate shifts that harm generalization Rosenfeld et al. (2020).

To formalize the utility–privacy trade-off, we represent the pooled data as a category (C), with samples as objects and admissible relations as morphisms. We then use functors to map this common structure into separate utility and privacy categories: (F_{util}) captures utility, while (F_{priv}) captures privacy Riehl (2017). This separation allows the same data and morphisms to be evaluated under distinct, potentially opposing objectives, making category theory a natural framework for studying their trade-off without collapsing them into a single score. The formulation is also extensible: additional objectives, such as fairness or robustness, can be introduced as new categories with corresponding functors from (C), (see Figure 1).

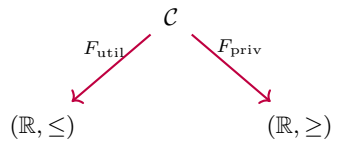


Figure 1: **Category theoretic lens.** Data samples belonging to category $\mathcal{C}$ can be mapped to categories $(\mathbb{R}, \leq)$ and $(\mathbb{R}, \geq)$ by the functors F_{util} and F_{priv} respectively.

Keeping the aforementioned points in mind, we propose a *privacy-aware pooling strategy* that integrates privacy into the matching-based harmonization framework: (i) a pooling specific (ε, δ) differential privacy metric (DP_{pool}); (ii) a privacy-tempered matching criterion that preferentially admits samples close to each institution's own population mean (those that incur the smallest DP_{pool} cost) without requiring access to private population statistics and using the matched core as a principled proxy; and (iii) a monotone radial scoring rule that jointly maximizes the number of admitted samples and their matching fidelity while enforcing a proposed pooling specific ϵ–δ privacy budget (DP_{pool}), so that privacy cost grows monotonically with inclusion. **Note.** Prior privacy-preserving pooling methods have focused on distributed statistical analysis through aggregation or virtual pooling of covariates in matched studies Saha-Chaudhuri & Weinberg (2017). In contrast,

our setting addresses privacy at the sample-selection stage itself, under heterogeneous domain shift and an explicit per-institution differential privacy budget. To the best of our knowledge, *this is the first matching-based pooling method that considers and proposes pooling-specific privacy at sample selection level and jointly handles harmonization, sample utilization, and explicit differential privacy budgeting under domain shift*. In the subsequent sections we discuss existing pooling techniques and their limitations (Sec. 2), describe our privacy-aware data pooling strategy (Sec. 3), and present experimental evidence against baselines (Sec. 4).

2 Preliminaries and Existing Pooling Techniques

Setup. Let $\mathcal{K} = \{1, \dots, K\}$ index a collection of institutions. Institution k has dataset $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}$, where $x_i^{(k)} \in \mathbb{R}^d$ is a feature vector and $y_i^{(k)} \in [C] := \{0, \dots, C-1\}$ is a class label. Let $\mathcal{D}_k^{(c)} = \{x \in \mathcal{D}_k : y = c\}$ be the class-c stratum of institution k. A *pooling strategy* selects a subset $\mathcal{P} \subseteq \bigcup_{k=1}^{K} \mathcal{D}_k$ used for downstream learning. Performance is measured relative to a *target distribution* $P^\star$ whose support is unknown.

Population statistics. For each institution k, let $\mu_k^{\text{all}} = \frac{1}{n_k} \sum_{x \in \mathcal{D}_k} x$, $\mu_k^{(c)} = \frac{1}{|\mathcal{D}_k^{(c)}|} \sum_{x \in \mathcal{D}_k^{(c)}} x$, and $\sigma_k^2 = \frac{1}{n_k} \sum_{x \in \mathcal{D}_k} \|x - \mu_k^{\text{all}}\|^2$. For any selected subset $X_k^{\text{sel}} \subseteq \mathcal{D}_k$ let $\mu(X_k^{\text{sel}})$ be its empirical mean.

Basic pooling strategies. The simplest strategy, $\mathcal{P}_{\text{naive}} = \bigcup_{k=1}^{K} \mathcal{D}_k$, maximizes sample count but is entirely blind to distributional heterogeneity. When label-inconsistent institutions are large (as typical in practice) their supervision signal dominates and degrades performance on $P^\star$ (Roy et al., 2026). Uniform subsampling to at most $\lfloor n_{\text{sub}}/C \rfloor$ samples per class per institution reduces size imbalance but does not remove corrupted supervision; it merely draws fewer incorrect labels. The theoretical cost of ignoring the source-target mismatch is well understood through the lens of density ratio estimation: the effective reweighting of source samples requires proper estimation whose quality determines the gap between empirical and target risk (Yu et al., 2026). When this ratio is poorly estimated (as is common under strong domain shift-importance) weighted objectives can exhibit systematic bias, rendering matching a practically superior alternative that directly constrains which samples enter the pool rather than reweighting all of them post-hoc.

Filtering-based pooling. Roy et al. (2026) introduces a geometric filtering criterion anchored at a reference institution a whose labeling convention is assumed to agree with $P^\star$ to effectively filter out confounding data. Per-class centroids are initialized from the anchor, $\hat{c}_c \leftarrow \frac{1}{|\mathcal{D}_a^{(c)}|} \sum_{x \in \mathcal{D}_a^{(c)}} x, c \in [C]$, and a candidate sample x from institution $k \neq a$ with label y is admitted if and only if

$$\|x - \hat{c}_y\| \;<\; \|x - \hat{c}_j\| \quad \forall\, j \neq y. \tag{1}$$

Centroids are updated online via an exponential moving average after each institution is processed. The matched subset $\mathcal{M}_k = \{x \in \mathcal{D}_k : \text{condition } 1\}$ provably filters label-inconsistent samples while retaining those geometrically aligned with $P^\star$. The matched core is $\mathcal{X}_{\text{core}} = \mathcal{D}_a \cup \bigcup_{k \neq a} \mathcal{M}_k$. This represents all the forms of filtering-based data pooling.

The research gap. Matching resolves the utility challenge decisively but is entirely *privacy-agnostic*: among the samples that pass Eq. 1, it imposes no preference for samples that are representative of the contributing institution's own population. For an unharmonized institution (one whose population mean μ_k^{all} lies far from the anchor, refer to Sec. B.1 for details) the subset selected by matching concentrates near the anchor's distribution, which is precisely the tail of institution k's own distribution. As we formalize in Sec. 3, this tail concentration leads to substantial privacy leakage even for samples that pass the matching filter.

3 Privacy-Aware Data Pooling from Multiple Sources

Definition 1 (**(ε, δ)-Differential Privacy**). *A randomised mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if, for every pair of neighbouring datasets D, D' and every measurable output set $\mathcal{O}$,*

$$\Pr[\mathcal{M}(D) \in \mathcal{O}] \;\leq\; e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in \mathcal{O}] + \delta.$$

Smaller ε is stronger; δ is a small failure probability (typically 10^{-6}–10^{-8}) bounding the measure of outcomes on which the guarantee may not hold. Which datasets count as *neighbouring* is a modelling

choice that fixes the unit being protected – one patient, one hospital, or something between – and multi-institutional pooling exposes three such units simultaneously (Sec. A.2). Given a statistic f with ℓ_2-sensitivity $S = \sup_{D \sim D'} \|f(D) - f(D')\|_2$, the Gaussian mechanism releases $f(D) + Z$, $Z \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I)$, and achieves (ε, δ)-DP with $\varepsilon = \frac{S\sqrt{2\ln(1.25/\delta)}}{\sigma_{\mathrm{DP}}}$.

Supremum is the wrong object at the pooling stage. ε is calibrated to a worst case over *all* datasets a site could conceivably hold — the right object when the dataset is given and the only lever is how much noise to add, but the wrong one here. First, it prices a hypothetical: the supremum is attained by a pathological institution, not the one in front of us, and inflating σ_{DP} to cover it destroys the supervision signal that data-scarce pooling exists to preserve (Müller et al., 2019; Rosenfeld et al., 2020). Second, it offers *no lever at all* on the decision we actually make: a pooling designer does not choose how noisy the release is, it chooses *which samples enter*. A privacy notion for pooling must therefore be a function of the selection which controls sample selection rather than the noise.

Definition 2 (**Pooling privacy cost,** $\mathrm{DP}_{\mathrm{pool}}$)**.** *Let institution k hold $\mathcal{D}_k$ with population mean μ_k^{all}, and let a pooling rule admit $X_k^{\mathrm{sel}} \subseteq \mathcal{D}_k$. Under the* in-domain *neighbourhood – D' replaces X_k^{sel} by the full $\mathcal{D}_k$ – the* realized *sensitivity of the released pooled mean with respect to site k is*

$$S_k \;=\; \left\|\mu(X_k^{\mathrm{sel}}) - \mu_k^{\mathrm{all}}\right\|_2, \qquad \varepsilon_k \;=\; \frac{S_k\sqrt{2\ln(1.25/\delta)}}{\sigma_{\mathrm{DP}}}. \tag{2}$$

A pooling design satisfies $(\varepsilon_k^{\max}, \delta)$-$\mathrm{DP}_{\mathrm{pool}}$ for site k if $\varepsilon_k \le \varepsilon_k^{\max}$ for the selection it returns.

Bi & Shen (2023) attack the same trade-off from the opposite side: it noises the data and transforms it back to retain the *source* distribution. The guarantee is incorrect for unharmonized institutions, where a confounded site's raw conclusion is exactly what pooling must discard, and Bi & Shen (2023) perturbs values while keeping every record, whereas pooling chooses which unmodified records to admit. Nor can the two be composed: privatizing each site before matching leaves S_k unchanged.

Three levels of privacy. What a pooling strategy protects is fixed by which datasets we call *neighbours*, and multi-institutional pooling exposes three units at once (formal definitions in Sec. A.2). **Sample level:** neighbors differ by one record, so the unit is a patient, and the cost of admitting a sample is its influence on the pooled mean. **In-domain level:** neighbors differ by replacing the contributed subset with the institution's full population, so the unit is the site's *distributional identity* (a selection drawn from a thin shell betrays structure about the population even when no record can be recovered). **Hospital level:** neighbors differ by an institution's entire contribution, so the unit is the fact of participation, and exposure compounds with both contribution size and distance from the rest of the pool. *The three are one attack at three granularities*.

We budget the in-domain level as it is the one that matters in continual pooling where institution data arrive incrementally and models are repeatedly updated. It also brackets the other two: sample-level exposure gives the intuition that central points are cheaper, while hospital-level exposure is inherited once in-domain sensitivity is controlled (Sec. A.2). [1]

3.1 From Matching's Privacy Failure to a Minimax Objective

The standard lens for analyzing privacy-utility trade-offs is statistical: utility is quantified via risk bounds or distributional alignment scores, privacy via differential privacy composition Dwork & Roth (2014), and their interplay through Lagrangian scalarizations or information-theoretic inequalities Zhong & Bu (2022); Padakandla et al. (2019). While powerful, this perspective treats utility and privacy as two scalar objectives to be traded off, collapsing the tension into a single composite score $\mathcal{L}(\lambda) = \mathcal{U}(x) - \lambda \cdot \mathcal{P}(x)$ ($\lambda \ge 0$). Such scalarization is sensitive to the choice of λ, non-convex in discrete selection problems Bertsekas (2014), and critically *obscures the structural origin* of the conflict. We adopt a category-theoretic viewpoint Riehl (2017) to reveal why this tension is not a numerical artifact but a *structural incompatibility* intrinsic to the sample space under domain shift.

A category of pooled samples. Recall each institution k contributes feature–label pairs $(x_i^{(k)}, y_i^{(k)})$ to the pooled candidate set $\mathcal{X} = \bigcup_k D_k$. For matching, a sample x from a con-

[1] The pooling protocol only releases the current pooled mean $\tilde{\mu}_k$ at stage k. Prior releases are not retained. External adversaries thus cannot form cross-stage differences $\tilde{\mu}_k - \tilde{\mu}_{k-1}$, and sequential composition over K site additions reduces to the per-stage $\mathrm{DP}_{\mathrm{pool}}$ guarantee, characterized empirically by the ε curves in Fig. 5 and Fig. 15. Participant adversaries retaining prior observed releases are out of scope.

founder hospital is desirable if it moves closer to the anchor's class centroid $\hat{c}_y$ and farther from the competing centroid $\hat{c}_{1-y}$, even when this motion pushes x into the tail of that hospital's own distribution. To capture such *motions* abstractly, we define a category $\mathcal{C}$ whose objects are samples $x \in \mathcal{X}$, and whose morphisms $f : x \to x'$ represent admissible transformations in feature space (e.g., small displacements along directions that improve class separation for the anchor). Composition is function composition, with identity morphisms id_x.

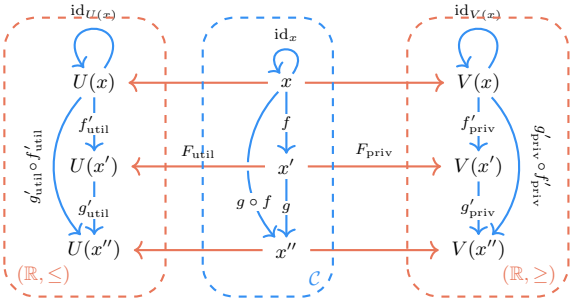


Figure 2: **Both functors on one diagram.** Admissible motions f, g in the sample category $\mathcal{C}$ are sent by F_{util} into $(\mathbb{R}, \leq)$ and by F_{priv} into $(\mathbb{R}, \geq)$, each preserving identities and composition, $F(g \circ f) = F(g) \circ F(f)$. The two targets are a poset and its opposite, which is why they exhaust the order-compatible possibilities: harmonization utility is covariant, in-domain privacy contravariant, and the conflict of Thm. 1 is this mismatch of variance.

Two functorial views: utility vs. privacy. For a fixed class y, let $\mathcal{C}_y$ denote the subcategory of admissible forward translations along the ray $x(r) = \mu_k^{\mathrm{all}} + r u_y$, $r \geq 0$, where $u_y := v_y / \|v_y\|_2$ and $v_y := \hat{c}_y - \hat{c}_{1-y} \neq 0$. We encode matching utility and in-domain privacy via two functors $F_{\mathrm{util}} : \mathcal{C}_y \to (\mathbb{R}, \leq)$ and $F_{\mathrm{priv}} : \mathcal{C}_y \to (\mathbb{R}, \geq)$ (see Fig. 2). We formalize the validity of these mappings and their monotone behavior in Assumption. 1 and Lemma. 4.

We define the object mappings as real-valued scores $U(x) = F_{\mathrm{util}}(x) = \|x - \hat{c}_{1-y}\|_2^2 - \|x - \hat{c}_y\|_2^2$ and $V(x) = F_{\mathrm{priv}}(x) = -\|x - \mu_k^{\mathrm{all}}\|_2^2$. The functors operate on morphisms by mapping $f : x \to x'$ to the unique target morphisms $f'_{\mathrm{util}} = F_{\mathrm{util}}(f)$ and $f'_{\mathrm{priv}} = F_{\mathrm{priv}}(f)$. Here larger $U(x)$ indicates better anchor alignment, while larger $V(x)$ indicates better in-domain privacy. Along morphisms in $\mathcal{C}_y$, the mapping to f'_{util} preserves order in $(\mathbb{R}, \leq)$ and the mapping to f'_{priv} preserves order in $(\mathbb{R}, \geq)$. $\mathrm{DP}_{\mathrm{pool}}$ is the selection-level form of the privacy functor: we can write $S_k^2 = -F_{\mathrm{priv}}\big(\mu(X_k^{\mathrm{sel}})\big)$, so ε_k is F_{priv} evaluated in the barycenter of the admitted set rather than in a single sample.

Geometric incompatibility. For a confounder institution Castro et al. (2020); Zhang et al. (2025); Zheng et al. (2025), the directions that improve matching utility and the directions that preserve in-domain privacy point are opposite. We model the incompatibility result on the forward-translation category $\mathcal{C}_y$ introduced above. Its objects are the points $x(r) = \mu_k^{\mathrm{all}} + r u_y$, $r \geq 0$, and its morphisms are translations $f_t : x(r) \mapsto x(r + t)$ for $t \geq 0$.

Theorem 1 (Functorial incompatibility under shift). *Let k be an unharmonized institution, and fix a class y such that $v_y := \hat{c}_y - \hat{c}_{1-y} \neq 0$. Then there exist composable morphisms $x \xrightarrow{f} x' \xrightarrow{g} x''$ in $\mathcal{C}_y$ such that F_{util} maps them to valid order-preserving morphisms f', g' in $(\mathbb{R}, \leq)$ satisfying $U(x) < U(x') < U(x'')$, whereas F_{priv} reverses the order, yielding $V(x) > V(x') > V(x'')$.*

Thm. 1 formalizes the key tension in Sec. 3: for unharmonized institutions (e.g., confounded hospital data), improving alignment with the anchor (mapping a sample to f' in utility space) necessarily pushes it toward the tail of its native distribution, increasing in-domain sensitivity and thus the DP cost of releasing the pooled mean. Prop. 2 ensures this trade-off is preserved after encoding: the encoder induces a map G through the matched-core pushout, so the composite $F \circ G$ evaluates the same anchor-consistent ordering in feature space as in Thm. 1. Because radial bands are defined from the same matched core M_k via $\hat{\mu}_k = \mu(M_k)$ (Stage 3), preserving the matching relation also preserves the core geometry underlying radial ordering. Therefore, Prop. 1 remains consistent with this anchor-certified structure in feature space.

Consequence for matching and toward a minimax objective. Matching selects samples that approximately maximize utility: the gate in 1 admits points whose motion under morphisms in $\mathcal{C}$ yields target morphisms f' in $(\mathbb{R}, \leq)$. By Thm. 1, following such morphisms from confounder institutions necessarily degrades privacy, i.e., moves samples away from μ_k^{all}. Combined with the radial-band ordering in Prop. 1, this implies that matching systematically over-selects outer-band (tail) samples and is, in this sense, structurally privacy-averse.

To reconcile utility and privacy, we view pooling for institution k as choosing a discrete subset $\mathcal{S}_k \subseteq \mathcal{D}_k$ of candidate samples and pose the following constrained problem:

$$\max_{\mathcal{S}_k} \frac{1}{|\mathcal{S}_k|} \sum_{x \in \mathcal{S}_k} U(x) + \log(|\mathcal{S}_k| + 1) \quad \text{s.t.} \quad \underbrace{\frac{\rho(\mathcal{S}_k)\sqrt{2\ln(1.25/\delta)}}{\sigma_{\text{DP}}}}_{\varepsilon_k} \leq \varepsilon_k^{\max} \tag{3}$$

where $\rho(\mathcal{S}_k) = \|\mu(\mathcal{S}_k) - \mu_k^{\text{all}}\|_2$ is the in-domain sensitivity (Def. 4) and $\varepsilon_k^{\max}$ is the institution's privacy budget. The log term encourages larger selections when privacy allows. Introducing a Lagrange multiplier $\lambda \geq 0$ Shapiro (1979) for the constraint yields the composite score

$$m_{\text{priv}}(x) = U(x) - \lambda \, \|x - \mu_k^{\text{all}}\|_2^2 \tag{4}$$

which can be viewed as an object-level composite score interpolating between F_{util} and F_{priv}.

3.2 Privacy-Budgeted Minimax Matching (PBMM)

PBMM is the concrete three-stage instantiation of this selector. Unlike a standard Lagrangian surrogate (Eq. 4), which collapses the privacy–utility trade-off into a multiplier-sensitive scalarization and can be brittle in discrete, nonconvex selection problems Shapiro (1979); Bertsekas (2014), PBMM preserves the minimax objective explicitly and implements it through a geometry-aware, budgeted admission rule that is fully data-driven and stage-wise.

3.2.1 Stage 1: Matched Core

As seen in Sec. 3.1, utility and privacy are high related but are conflicting objectives. Thus, we start by maintaining utility in Stage 1 by utilizing the matching gate (Eq. 1) Stuart (2010); Roy et al. (2026); Fan et al. (2025), applied across all institutions to obtain matched subsets $\mathcal{M}_k$ and a core pool $X_{\text{core}} = D_a \cup \bigcup_{k \neq a} \mathcal{M}_k$. Institutions with $\mathcal{M}_k = \emptyset$ are *hard-excluded*: without any matched slice, these hospitals have no objects in the well-aligned region of utility space and so cannot be reached by paths that preserve matching structure, and any *privacy correction* would collapse back towards naive pooling (see Sec. 3.1). For institutions with $|\mathcal{M}_k| > 0$, the matched core provides: (i) a causally trusted subset for each hospital (ii) consistent anchor centroids. Their empirical means define $\hat{\mu}_k := \mu(\mathcal{M}_k), \quad \hat{\sigma}_k^2 := \frac{1}{|\mathcal{M}_k|} \sum_{x \in \mathcal{M}_k} \|x - \hat{\mu}_k\|_2^2$, which summarize the in-domain geometry of the admitted slice and will serve as our privacy reference in Stage 2.

3.2.2 Stage 2: Privacy-Tempered Expansion

After Stage 1, each institution k with $|\mathcal{M}_k| > 0$ retains $n_k - |\mathcal{M}_k|$ left-out samples. Eq. 4 tells us that, along any admissible path in feature/data space, we should prefer samples that: (i) increase F_{util} (ii) stay close to the institution's population mean to control F_{priv}. The role of Stage 2 is to bias the margin toward better DP_{pool} (F_{util}), using only statistics computed from $\mathcal{M}_k$ to address the tradeoff in Theorem 1 as we do not assume access to full-dataset quantities such as μ_k^{all} during pooling [2]

Instead, when $M_k > 0$ and M_k is large, this signifies nontrivial overlap with existing matched set and the added domain's distribution, so $\hat{\mu}_k$ serves as the closest available proxy ($\hat{\mu}_k \sim \mu_k^{\text{all}}$) as shown in Lemma 1 (see Sec. C.1 for proof). By Lemma 1, a nontrivial matched core M_k certifies sufficient overlap with the anchor-consistent region, so $\hat{\mu}_k = \mu(M_k)$ is the closest available proxy for μ_k^{all} under our pooling protocol. If $M_k = \varnothing$, the institution is hard-excluded already in Stage 1.

Lemma 1 (Matched-core mean as a proxy under sufficient overlap). *Let $D_k = M_k \sqcup L_k$ with $\alpha_k := |M_k|/|D_k| \in (0, 1]$, $\hat{\mu}_k := \mu(M_k)$, $\mu_k^{\text{all}} := \mu(D_k)$, and $\mu_{L,k} := \mu(L_k)$. Assume institution k is τ_k-overlapping with the anchor-consistent region in the sense that $\|\hat{\mu}_k - \mu_{L,k}\|_2 \leq \tau_k$. Then $\|\hat{\mu}_k - \mu_k^{\text{all}}\|_2 \leq (1 - \alpha_k)\tau_k$. In particular, if matching returns a substantial slice (α_k not small) and the matched slice lies in a local overlap region of radius τ_k, then $\hat{\mu}_k$ is a good proxy for the full institution mean μ_k^{all}.*

[2] This design enforces a clear separation of roles. During pooling, the central aggregator observes only incrementally admitted samples and cannot access μ_k^{all}, which remains private to institution k. Sensitivity computation (Algorithm 1) is performed *locally*, and only ε_k is shared—never μ_k^{all} itself. This follows a standard federated protocol Aledhari et al. (2020): the aggregator enforces the global budget using reported ε_k, while institutions retain full control over private statistics. Consequently, all geometric computations during pooling use $\hat{\mu}_k$ as a proxy, and μ_k^{all} is used solely for local budget verification.

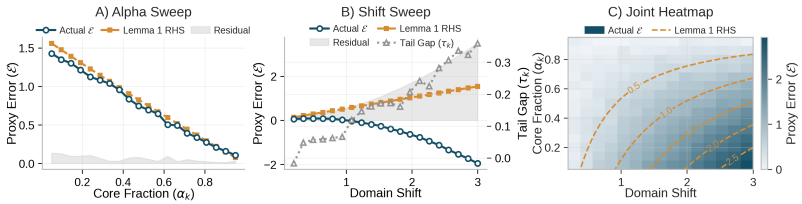


Figure 3: **Empirical validation of Lemma 1.** Each subplot isolates a governing variable in the bound $|\hat{\mu}_k - \mu_k^{\text{all}}|_2 \leq (1-\alpha_k)\tau_k$, which justifies using $\hat{\mu}_k$ as a privacy-safe proxy for the inaccessible population mean μ_k^{all} in Stage 2. **(A) Alpha Sweep:** With fixed domain shift, varying the matched core M_k shows a strictly linear decrease in proxy error with α_k, exactly matching the bound. **(B) Shift Sweep:** With fixed α_k, increasing domain shift increases the tail gap τ_k, and the proxy error scales proportionally, yet remains controlled whenever a non-empty core exists. **(C) Joint Surface:** Over the full (α_k, shift) grid, empirical error closely follows $(1-\alpha_k)\tau_k$ with negligible residuals.

We empirically validate that $\hat{\mu}_k$ is a reliable proxy for μ_k^{all}. If this proxy were inaccurate, m_{priv} (Eq. 4) would lose geometric meaning, and the DP-cost ordering from Proposition 1 would no longer reflect true sensitivity. Lemma 1 bounds the error as $|\hat{\mu}_k - \mu_k^{\text{all}}|_2 \leq (1-\alpha_k)\tau_k$, where $\alpha_k = |M_k|/|D_k|$ (core fraction) and τ_k is the core–tail gap. This bound becomes tight when matching succeeds: a large, consistent M_k ensures both alignment (utility) and proximity to μ_k^{all} (privacy). We isolate these factors through three ablations (Sec. D.2). Varying α_k with fixed shift (Fig. 3A), the error decreases linearly, matching the bound and confirming Stage 1's exclusion behavior (domains with $M_k = \emptyset$ exhibit maximal error). Fixing α_k and increasing shift (Fig. 3B), the error scales with τ_k, showing stability even under strong domain mismatch. Finally, joint evaluation over (α_k, shift) (Fig. 3C) closely follows the theoretical surface $(1-\alpha_k)\tau_k$, indicating the bound is effectively tight. Overall, the matched-core mean is a geometry-certified proxy: reliable when matching conditions hold.

Privacy-tempered centroids and margin. We introduce a privacy-tempered own-class centroid for institution k, computed by reweighting its full domain around the core mean $\hat{\mu}_k$:

$$\tilde{c}_y^{(k)} = \frac{\sum_{x \in \mathcal{D}_k^{(y)}} w_k(x)\, x}{\sum_{x \in \mathcal{D}_k^{(y)}} w_k(x)}, \quad w_k(x) = \exp\left(-\frac{\|x-\hat{\mu}_k\|_2^2}{2\hat{\sigma}_k^2}\right) \tag{5}$$

The Gaussian kernel Wand & Schucany (1990) is used as the default smooth radial weighting around $\hat{\mu}_k$. Under our protocol, M_k is the available anchor-consistent slice of institution k. For institutions with nontrivial overlap, as $|M_k|$ grows, the empirical center and spread $(\hat{\mu}_k, \hat{\sigma}_k^2)$ stabilize by the law of large numbers Vershynin (2018). Thus weighting by $w_k(x) = \exp\big(-\|x-\hat{\mu}_k\|_2^2/(2\hat{\sigma}_k^2)\big)$ naturally gives highest influence to samples from the typical in-domain region while smoothly down-weighting tail points De Alvis & Seneviratne (2024), rather than imposing a hard cutoff. The resulting centroid $\tilde{c}_y^{(k)}$ therefore pulls the own-class reference back toward the in-domain core while leaving the anchor centroids $\hat{c}_y$ unchanged. Sec. D.2 shows that alternative radial kernels are also possible, but Gaussian and Laplacian give the strongest overall trade-off.

$$m_{\text{priv}}(x) = \frac{1}{R_k^2}\Big[\|x-\hat{c}_{1-y}\|_2^2 - \|x-\tilde{c}_y^{(k)}\|_2^2\Big], \quad R_k = \max_{x' \in \mathcal{D}_k \setminus \mathcal{M}_k} \|x'-\hat{\mu}_k\|_2 \tag{6}$$

Eq. 6 is the privacy-aware geometric instantiation of the Lagrangian trade-off in 4 expressed in terms of our two functors: $m_{\text{priv}}(x) \propto F_{\text{util}}(x) + F_{\text{priv}}(x)$. Using $\tilde{c}_y^{(k)}$ and $\tilde{c}_{1-y}^{(k)}$ moves selection closer to $\hat{\mu}_k \sim \mu_k^{all}$, inadvertently making samples closer to μ_k^{all} easier to admit. [3]

3.2.3 Stage 3: Radial Bucket Scoring and Monotone Expansion

Let $X_k^{\text{sel}} = \mathcal{M}_k$ initially. Among left-out samples, those with $m_{\text{priv}}(x) > 0$ are admissible under the minimax criterion Boyd & Vandenberghe (2004). For each participating institution k, we partition admissible samples into three concentric radial bands by normalized distance to $\hat{\mu}_k$:

$$\mathcal{B}_b^{(k)} = \left\{x \in \mathcal{D}_k \setminus \mathcal{M}_k : \frac{\|x-\hat{\mu}_k\|_2}{R_k} \in I_b,\ m_{\text{priv}}(x) > 0\right\} \tag{7}$$

with $I_{\text{in}} = [0, \frac{1}{3})$, $I_{\text{mid}} = [\frac{1}{3}, \frac{2}{3})$, $I_{\text{out}} = [\frac{2}{3}, 1]$. By Proposition 1, these bands satisfy $\varepsilon_k(\mathcal{B}_{\text{in}}) \leq \varepsilon_k(\mathcal{B}_{\text{mid}}) \leq \varepsilon_k(\mathcal{B}_{\text{out}})$: inner bands are always cheapest in terms of in-domain DP cost. *We use $\hat{\mu}_k$ for*

[3] When $\hat{\sigma}_k^2 \to 0$, weights concentrate on $\mathcal{M}_k$ and $\tilde{c}_y^{(k)} \to \hat{c}_y$, so $m_{\text{priv}}(x)$ reduces to the matching margin $m_{\text{match}}(x)$. When $\hat{\sigma}_k^2 \to \infty$, all weights are equal and $\tilde{c}_y^{(k)}$ becomes the unweighted class mean; with no score threshold, all samples are admitted, recovering naive pooling. Thus PBMM interpolates continuously between the two extremes (Sec. 3.1), with $\hat{\sigma}_k^2$ a data-driven *privacy temperature* determined entirely by the matched core.

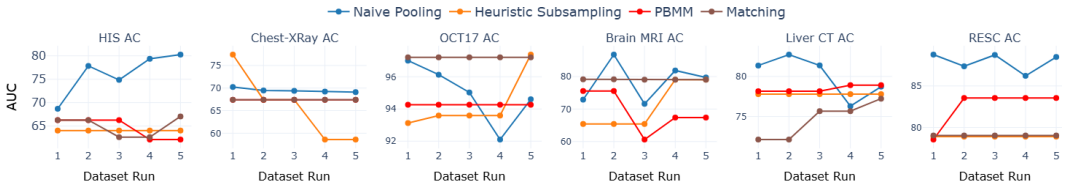


Figure 4: **Downstream AUC across six AC medical imaging benchmarks over five cumulative dataset addition runs, comparing Naive Pooling, Heuristic Subsampling, Matching, and PBMM (ours).** DA scores (higher is better) for AC settings are as follows—Naive Pooling: HIS 3.26, Chest-XRay 0.00, OCT17 1.10, Brain MRI 2.45, Liver CT 2.11, RESC 2.12; Heuristic Subsampling: HIS 4.00, Chest-XRay 2.00, OCT17 4.16, Brain MRI 4.41, Liver CT 4.00, RESC 4.00; Matching: HIS 3.18, Chest-XRay 4.00, OCT17 4.00, Brain MRI 3.00, Liver CT 4.13, RESC 4.00; PBMM: HIS 3.00, Chest-XRay 4.00, OCT17 4.00, Brain MRI 3.20, Liver CT 4.02, RESC 4.05. PBMM achieves DA scores competitive with Matching across all six benchmarks (same on Chest-XRay, OCT17, and RESC AC) while outperforming Naive Pooling substantially (e.g., Chest-XRay: 4.00 vs. 0.00) and remaining broadly comparable to Heuristic Subsampling. Crucially, PBMM attains stronger or comparable final AUC than Matching in several settings (e.g., RESC, Brain MRI) while enforcing a principled in-domain privacy budget during sample selection (Fig. 15).

the bands I_b as $\hat{\mu}_k \sim \mu_k^{all}$ (see Stage 2) which indicates I_b approximately mimics the actual bands of $\mathcal{D}_k$. This indicates that the true minmax objective (Eq. 3) is respected. The radial bands are therefore formed only for institutions with $|M_k| > 0$, since only those institutions contribute a certified core whose geometry can be expanded monotonically under the privacy budget.

Each band $\mathcal{B} = \mathcal{B}_b^{(k)}$ receives a scalar score 8 (Sec. D.2 provides detailed ablation) combining sample count (with diminishing returns) and average privacy-tempered margin, where $\log(|\mathcal{B}| + 1)$ captures the marginal gain from adding more samples already present in the band, while the mean m_{priv} penalizes bands whose additional samples lie farther from the institution's core.

$$\text{score}(\mathcal{B}) = \log(|\mathcal{B}| + 1) + \tfrac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} m_{\text{priv}}(x) \tag{8}$$

PBMM then performs *monotone radial expansion*: at each step it admits the highest-scoring band subject to a positive score and to radial order within each institution, inner before mid before outer. That ordering is the operational form of Prop. 1 (a hospital exhausts its cheap central bands before contributing tail samples) and it makes the procedure a direct, *hyperparameter-free* approximation to equation 3, with no Lagrangian to solve. Each institution operates under an $(\varepsilon_k^{\max}, \delta)$-$\text{DP}_{\text{pool}}$ budget: admitting a band updates its in-domain sensitivity, and if the resulting ε_k would exceed $\varepsilon_k^{\max}$ the band's score is set to $-\infty$. Sensitivity is evaluated *cumulatively*, $S_k = \|\mu(X_k^{\text{sel}} \cup \mathcal{B}_b^{(k)}) - \mu_k^{\text{all}}\|_2$ with X_k^{sel} already containing every previously admitted band, so it measures the true marginal cost rather than the band's cost in isolation. The expansion rule is otherwise unchanged, and Lem. 3 shows the enforcement is a loop invariant, so the released $\mu(\mathcal{P})$ satisfies the budget.

4 Experimental Results

Setups. We evaluate under four progressively challenging settings, each targeting a distinct aspect of the privacy–harmonization tension (Sec. 3). **i) Synthetic Gaussian:** A controlled 2D Gaussian setting with known ground truth, matching our theoretical assumptions, used to empirically validate the privacy–utility trade-off. **ii) Homogeneous real data:** Autism Brain Imaging Data Exchange (ABIDE) Nielsen et al. (2013), where sites share modality and labels but differ in scanners and acquisition protocols, is used. This common "homogeneous shift" scenario tests effectiveness in high-dimensional yet unimodal feature spaces Shen et al. (2024). **iii) Heterogeneous real data:** The most challenging setting pools across fundamentally different imaging modalities in a zero-shot regime Roy et al. (2026); Huang et al. (2024) operating in the non-Euclidean embedding space of CLIP Radford et al. (2021). Prop. 2 ensures anchor-consistent ordering is preserved from data to feature space when the matching criterion

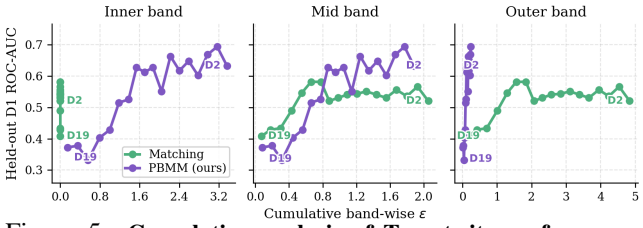


Figure 5: **Cumulative analysis of Target-site performance across privacy bands for each domain addition in the homogeneous ABIDE setting.** We report held-out D1 ROC AUC as a function of cumulative band-wise ϵ.

holds. **iv) Pooling under data scarce scenario:** We perform sequential data additions from one cohort to another to study robustness to distribution shift as external data becomes available to a small anchor cohort Shen et al. (2024). We start from a 43-subject in-house cohort and sequentially add participants from ADNI Petersen et al. (2010) and evaluate age-group classification from functional brain connectomes. **Remark.** The in-house dataset contains anonymized fMRI scans from 43 participants aged 12-73 years. To protect participant privacy, these data are not publicly disclosed during peer review. Its limited sample size provides a realistic data-scarce setting in which we evaluate whether selective pooling with public ADNI data can increase effective sample support while limiting performance degradation under cross-cohort distribution shift. Please see Sec. B.1,B.2 and B.3 for additional details.

We primarily use two evaluation metrics: i) DP_{pool} (Def. 2) for privacy ii) DA score for utility. Following Roy et al. (2026), pooling is judged by *non-deterioration* across sequential domain additions rather than by accuracy at one checkpoint: given held-out performance $y_1, \ldots, y_{T+1}$ after each of T additions, $\mathrm{DA} = \sum_{i=1}^{T} \mathbb{1}\{y_{i+1} \geq y_i\}\left(1 + \frac{y_{i+1}-y_i}{10} s_i\right)$ with weights s_i increasing in i, so a step that degrades performance scores zero and $\mathrm{DA} \approx T$ indicates no deterioration.

Empirical Validation. Extensive experimental study of the proposed method compared to the existing commonly used pooling techniques for the synthetic Gaussian setup and the homogeneous setup justify the theoretical picture of the utility-privacy tradeoff and PBMM achieves better privacy (see Fig. 5) while maintaining comparable utility (See Table 1). As seen in Fig. 5, PBMM achieves stronger overall privacy-utility trade-off across all three bands. Matching remains competitive only when enough well-aligned samples are available, but tends to over-filter and under-utilize available data. Fig. 6 highlights the results for the data scarce setting where matching provides the most stale trajectory and the highest stage-averaged accuracy. PBMM on the other hand admits more cross-cohort samples and recovers strongly as the external pool grows while matching suffers from overfiltering. The results highlight that adding out-of-distribution data can degrade performance in data-scarce settings, while selective pooling controls this degradation. For heterogeneous setting, the modality cones are non overlapping Roy et al. (2026) adding to the already difficult zero-shot high heterogeneity. Thus matching shows its maximum potential where it potentially reduces the gap. *This aligns with modality gap vs performance phenomena observed in Liang et al. (2022); Levi & Gilboa (2024).* We follow the same experimental setup as Roy et al. (2026) (See Sec. B.3 for further details). It can be seen in Fig. 4 that PBMM performs comparable to existing pooling techniques in terms of DA score and even final performance demonstrating the ability to harmonize pooled sets to prevent performance dips. Fig. 15 shows that PBMM performs better overall across across all privacy bands (inner, mid, outer). Since matching is a filtering based pooling technique, it over filters samples which leads to under-utilization of available resource as seen in Chest-XRay. Additional experiments can be seen in Sec. D.2, A, B.2, B.3, B.1, B.4.

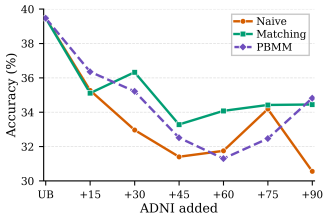


Figure 6: **Pooled distribution vs. Test distribution.** Filtering-based pooling techniques (matching and PBMM) outperform traditional techniques. Average performance reported for stratified three-fold cross-validation over six random seeds.

5 Conclusion

We study multi-institutional data pooling under distribution shift, where harmonization through sample selection can improve utility but also introduce privacy leakage at the point of selection. To address this, we introduce the pooling-specific (DP_{pool}) notion and use a category-theoretic formulation to characterize the structural trade-off between utility and privacy. Building on this view, we propose PBMM, a privacy-budgeted matching framework that jointly promotes harmonization, sample utilization, and privacy preservation. Our preliminary data-addition experiments further indicate that performance continues to improve as more samples are incorporated while the proposed privacy guarantees are maintained. We also evaluate a practical setting using an in-house dataset, referred to as such because it must remain anonymized during peer review, and harmonize it with a public dataset to expand the effective sample size. This experiment represents an early step toward using privacy-aware harmonization to increase the statistical power and broader utility of institution-specific datasets.

Table 1: **Homogeneous ABIDE results on the held-out target site D1 (Fig. 14).** Best result per column is **bolded**. DA score is the utility metric for pooling stage evaluation, following the continual-pooling protocol of Roy et al. (2026).

Method	Acc.	DA Score
Naïve Pooling	57.14	10.37
Matching	60.70	**12.11**
PBMM (ours)	**61.71**	11.31

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Uzair Akbar, Niki Kilbertus, Hao Shen, Krikamol Muandet, and Bo Dai. An analysis of causal effect estimation using outcome invariant data augmentation. *arXiv preprint arXiv:2510.25128*, 2025.

Mohammed Aledhari, Rehma Razzak, Reza M Parizi, and Fahad Saeed. Federated learning: A survey on enabling technologies, protocols, and applications. *IEEE access*, 8:140699–140725, 2020.

James Bailie, Ruobin Gong, and Xiao-Li Meng. Differential privacy meets invariant statistics: Some conundrums in quantifying trade-offs. Technical report, National Bureau of Economic Research, 2024.

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4042–4053, 2024.

Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

Xuan Bi and Xiaotong Shen. Distribution-invariant differential privacy. *Journal of econometrics*, 235 (2):444–453, 2023.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Sotirios Panagiotis Chytas, Vishnu Suresh Lokhande, Peiran Li, and Vikas Singh. Pooling image datasets with multiple covariate shift and imbalance. *arXiv preprint arXiv:2403.02598*, 2024.

Charika De Alvis and Suranga Seneviratne. A survey of deep long-tail classification advancements. *arXiv preprint arXiv:2404.15593*, 2024.

Cynthia Dwork. Differential privacy. In *Encyclopedia of Cryptography, Security and Privacy*, pp. 649–652. Springer, 2025.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Jikang Fan, Shintaro Terabe, and Hideki Yaginuma. Propensity score matching–difference-in-differences analysis of the casual effect of opening intermediate high-speed railway stations on employment status in surrounding municipalities. *Case Studies on Transport Policy*, pp. 101592, 2025.

Sarang Galada, Tanurima Halder, Kunal Deo, Ram P Krish, and Kshitij Jadhav. Prism: Privacy-preserving inter-site mri harmonization via disentangled representation learning. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2025.

Quan Geng, Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184, 2015.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Alexander Hägele, Aryo Pradipta Gema, Henry Sleight, Ethan Perez, and Jascha Sohl-Dickstein. The hot mess of ai: How does misalignment scale with model intelligence and task complexity? *arXiv preprint arXiv:2601.23045*, 2026.

Jihun Hamm. Preserving privacy of continuous high-dimensional data with minimax filters. In *Artificial Intelligence and Statistics*, pp. 324–332. PMLR, 2015.

Ali Hatamizadeh, Hongxu Yin, Holger R Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10021–10030, 2022.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11375–11385, 2024.

Jane Kaye. The tension between data sharing and the protection of privacy in genomics research. *Annual review of genomics and human genetics*, 13(1):415–431, 2012.

Salome Kazeminia, Carsten Marr, and Bastian Rieck. Topological inductive bias fosters multiple instance learning in data-scarce scenarios. *arXiv preprint arXiv:2307.14025*, 2023.

Minjun kim, Dong Ju Mun, Hwihun Jeong, Haechang Lee, Se Young Chun, and Jongho Lee. Privacy-preserving MRI data harmonization for black-box models, 2025. URL `https://openreview.net/forum?id=NnsuEaN1Wx`.

Yongnam Kim and Peter Steiner. Quasi-experimental designs for causal inference. *Educational psychologist*, 51(3-4):395–405, 2016.

Alistair Lawley, Rory Hampson, Kevin Worrall, and Gordon Dobie. A cost focused framework for optimizing collection and annotation of ultrasound datasets. *Biomedical Signal Processing and Control*, 92:106048, 2024.

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of clip. *arXiv preprint arXiv:2411.14517*, 2024.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Vishnu Suresh Lokhande, Songwong Tasneeyapant, Abhay Venkatesh, Sathya N Ravi, and Vikas Singh. Generating accurate pseudo-labels in semi-supervised learning and avoiding overconfident predictions via hermite polynomial activations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11435–11443, 2020.

Vishnu Suresh Lokhande, Rudrasis Chakraborty, Sathya N Ravi, and Vikas Singh. Equivariance allows handling multiple nuisance variables when analyzing pooled neuroimaging datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10432–10441, 2022.

You-Wei Luo and Chuan-Xian Ren. When invariant representation learning meets label shift: Insufficiency and theoretical insights. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9407–9422, 2024.

Saunders Mac Lane and Saunders MacLane. *Categories for the working mathematician*, volume 5. Springer, 1971.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pp. 7313–7324. PMLR, 2021.

Michael Mattioli. The data-pooling problem. *Berkeley Technology Law Journal*, 32(1):179–236, 2017.

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.

Raghav Mehta, Changjian Shui, and Tal Arbel. Evaluating the fairness of deep learning uncertainty estimates in medical image analysis. In *Medical Imaging with Deep Learning*, pp. 1453–1492. PMLR, 2024.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31, 2018.

Daniel Moyer, Greg Ver Steeg, Chantal MW Tax, and Paul M Thompson. Scanner invariant representations for diffusion mri harmonization. *Magnetic resonance in medicine*, 84(4):2174–2189, 2020.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. Visualizing privacy-utility trade-offs in differentially private data releases. *arXiv preprint arXiv:2201.05964*, 2022.

Zeinab Navidi, Akshaya Thoutam, Madeline Hughes, Srivatsan Raghavan, Peter S Winter, Lorin Crawford, and Ava P Amini. Adaptive resampling for improved machine learning in imbalanced single-cell datasets. *bioRxiv*, pp. 2025–11, 2025.

Jared A Nielsen, Brandon A Zielinski, P Thomas Fletcher, Andrew L Alexander, Nicholas Lange, Erin D Bigler, Janet E Lainhart, and Jeffrey S Anderson. Multisite functional connectivity mri classification of autism: Abide results. *Frontiers in human neuroscience*, 7:599, 2013.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Arun Padakandla, PR Kumar, and Wojciech Szpankowski. The trade-off between privacy and fidelity via ehrhart theory. *IEEE Transactions on Information Theory*, 66(4):2549–2569, 2019.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3): 201–209, 2010.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Emily Riehl. *Category theory in context*. Courier Dover Publications, 2017.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Ayush Roy, Samin Enam, Jun Xia, Vishnu Suresh Lokhande, and Won Hwa Kim. Is exchangeability better than iid to handle data distribution shifts while pooling data for data-scarce medical image segmentation? *arXiv preprint arXiv:2507.19575*, 2025.

Ayush Roy, Rudrasis Chakraborty, Lav Varshney, and Vishnu Suresh Lokhande. Beyond pooling: Matching for robust generalization under data heterogeneity. *arXiv preprint arXiv:2602.07154*, 2026.

P Saha-Chaudhuri and CR Weinberg. Addressing data privacy in matched studies via virtual pooling. *BMC medical research methodology*, 17(1):136, 2017.

Jeremy F Shapiro. A survey of lagrangean techniques for discrete optimization. In *Annals of Discrete Mathematics*, volume 5, pp. 113–138. Elsevier, 1979.

Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y Chen. The data addition dilemma. *arXiv preprint arXiv:2408.04154*, 2024.

Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime Carbonell, and Kun Zhang. Domain adaptation with invariant representation learning: What transformations to learn? *Advances in neural information processing systems*, 34:24791–24803, 2021.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Matthew P Wand and William R Schucany. Gaussian-based kernels. *Canadian Journal of Statistics*, 18(3):197–204, 1990.

Yu-Xiang Wang. Per-instance differential privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019.

Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. *arXiv preprint arXiv:2511.13421*, 2025.

Mengyue Yang, Zhen Fang, Yonggang Zhang, Yali Du, Furui Liu, Jean-Francois Ton, Jianhong Wang, and Jun Wang. Invariant learning via probability of sufficient and necessary causes. *Advances in Neural Information Processing Systems*, 36:79832–79857, 2023.

Chenglin Yu, Zhengyu Zhou, and Weiwei Liu. Quantitative analysis of the effect of density ratio estima-tion in covariate shift adaptation. *Trans. Mach. Learn. Res.*, 2026, 2026.

Guixian Zhang, Guan Yuan, Debo Cheng, Lin Liu, Jiuyong Li, Ziqi Xu, and Shichao Zhang. Deconfounding representation learning for mitigating latent confounding effects in recommendation: G. zhang et al. *Knowledge and Information Systems*, 67(7):5999–6020, 2025.

Jiajing Zheng, Alexander D'Amour, and Alexander Franks. Copula-based sensitivity analysis for multi-treatment causal inference with unobserved confounding. *Journal of Machine Learning Research*, 26(36):1–60, 2025.

Hao Zhong and Kaifeng Bu. Privacy-utility trade-off. *arXiv preprint arXiv:2204.12057*, 2022.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

APPENDIX

A PRIVACY DEFINITIONS FOR DATA POOLING

A.1 DIFFERENTIAL PRIVACY IN MULTI-INSTITUTIONAL POOLING

What S_k means. It measures *how unrepresentative the contributed subset is of the site that contributed it*. If $S_k \approx 0$ the released mean is what an i.i.d. sample from the site would have produced, so an adversary differencing pooled means (Sec. 3) learns nothing beyond the site's ordinary statistics. If S_k is large the selection was drawn from a thin shell of the population, and the difference in means reports the location of that shell – revealing that the site harbours an unusual subgroup, a rare-condition cluster, or a systematic demographic skew, without any individual record being recovered. Two properties follow immediately and are worth stating because they are what make S_k the right quantity here rather than a convenient one. It is *shift-invariant*: translating an entire institution leaves S_k unchanged, so a hospital is never charged for being different from the consortium, only for misrepresenting itself. And it is *selection-controllable*: unlike the global supremum, it moves when the admission rule moves, which is precisely the lever PBMM operates.

Example 2 (DP_{pool} **Budget).** *Take $\sigma_{\mathrm{DP}} = 0.25$ and $\delta = 10^{-6}$, so $\sqrt{2\ln(1.25/\delta)}/\sigma_{\mathrm{DP}} = 21.2$ and $\varepsilon_k = 21.2 \cdot S_k$. Consider a confounder site from Sec. B.1 with in-domain radius $R_k = 3.0$. A* representative *contribution, e.g. a random stratified draw, gives $S_k \approx 0.05$ and $\varepsilon_k \approx 1.1$: the release is statistically indistinguishable from an ordinary sample of the site.* Matching *admits only samples near the anchor, which for a shifted site are its outermost samples; on $\mathcal{D}_5$ this yields $S_k = 2.40$ and $\varepsilon_k = 50.9$ (Fig. 8). The contributed mean sits four-fifths of the way to the edge of the site's own support, and saying so is itself the leak.*

We state the scope plainly, because S_k is a realized rather than worst-case sensitivity and the distinction matters. $\mathrm{DP}_{\mathrm{pool}}$ *is* an instance-specific accounting of the exposure created by the selection actually made, in the spirit of per-instance differential privacy (Wang, 2019) and of the data-dependent sensitivity used in smooth-sensitivity analyses (Nissim et al., 2007); it is a geometric budget on how atypical an admitted subset may be. It is *not* a worst-case algorithmic guarantee in the sense of DP-SGD (Abadi et al., 2016), and we do not claim end-to-end privacy of the trained model. Two consequences are worth being explicit about. Because S_k is a realized value, it is dominated by the worst-case in-domain sensitivity, so ε_k *lower-bounds* the ε that Definition 1 would assign to the same release: our numbers are the optimistic ones, and we report them as such. And because S_k is computed from μ_k^{all}, it is evaluated locally by each site; only the scalar ε_k is communicated, never the population mean itself (Sec. 3.2.2).

Upgrading to a worst-case guarantee. The gap above is closable, and the closure is a design choice rather than a new analysis. Suppose the admission radius is capped at an *a priori* constant r instead of the data-dependent R_k of Eq. 7: every admitted sample lies within r of the core mean, so every selection the rule can *ever* return satisfies $S_k \le r$, the supremum in Definition 1 is attained by construction, and releasing $\mu(\mathcal{P}) + \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I)$ satisfies $(r\sqrt{2\ln(1.25/\delta)}/\sigma_{\mathrm{DP}},\ \delta)$-DP in the standard sense under the in-domain neighbourhood. The price is adaptivity: a fixed r cannot respond to sites of differing spread, and in our experiments the adaptive radius admits substantially more samples at equal ε. We therefore report $\mathrm{DP}_{\mathrm{pool}}$ throughout and note that the worst-case variant is available at a known cost.

Pooled mean is the released statistic. Centroid-based harmonization, matching included (Roy et al., 2026), operates entirely through empirical means and Euclidean distances, so $\mu(\mathcal{P})$ is the sufficient statistic of the procedure for Gaussian and more general elliptical families (Vershynin, 2018). DP training can itself be viewed as repeatedly releasing means of clipped per-example gradients (Abadi et al., 2016), whose ℓ_2 sensitivity is governed by the same geometry, so tracking the pooled mean is a model-agnostic proxy for downstream cost. And the pooled mean is the object the aggregator forms directly from each site's admitted contribution, so the exposure exists at selection time rather than being created by training; $\mathrm{DP}_{\mathrm{pool}}$ budgets exactly the quantity that is exposed. $\mathrm{DP}_{\mathrm{pool}}$ budgets exactly the quantity that is exposed.

Example 3 (**Atypical records dominate**). *Consider a membership adversary who suspects record x_i from site k entered the pool and tests the shift $\|\mu(\mathcal{P}) - \mu(\mathcal{P} \setminus \{x_i\})\|_2$. Globally this scales as $\|x_i - \mu(X_k^{\mathrm{sel}})\|_2/(N-1)$ with $N = |\mathcal{P}|$; within site k's own contribution of size n it scales as $\|x_i - \mu(X_k^{\mathrm{sel}})\|_2/(n-1)$ (Def. 3), larger by a factor N/n. Concretely, for a site contributing $n = 200$ samples with centered mean and unit variance, a typical patient at distance 0.3 shifts the mean by $0.3/199 \approx 0.0015$, negligible against DP noise, whereas an atypical patient at distance 4.2 shifts it by $4.2/199 \approx 0.021$ – more than an order of magnitude larger, and detectable at $\sigma_{\mathrm{DP}} = 0.25$ after few repetitions. Atypical records are thus individually dangerous even when amortized across a large pool, and a selection rule that concentrates on them compounds the effect across every record it admits. This is the per-record shadow of the distributional statement in Example 2.*

Lemma 2 (**Mean-shift leakage is membership inference**). *Let the release be $\tilde{\mu} = \mu(\mathcal{P}) + Z$ with $Z \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I)$, and let an adversary decide between two neighbouring worlds whose released means differ by Δ. The optimal such test has advantage*

$$\mathrm{adv}(\Delta) \;=\; \mathrm{TPR} - \mathrm{FPR} \;=\; 2\,\Phi_{\mathcal{N}}\!\Big(\tfrac{\|\Delta\|_2}{2\sigma_{\mathrm{DP}}}\Big) - 1, \tag{9}$$

strictly increasing in $\|\Delta\|_2$, where $\Phi_{\mathcal{N}}$ is the standard normal CDF. Instantiating Δ at the three relations of Sec. A.2 gives $\|\Delta\|_2 = \Delta_{\mathrm{samp}}(x_i)$, S_k and $\Delta_{\mathrm{hosp}}(k)$ respectively. In particular the sample-level case is the classical membership-inference adversary, and since Def. 2 sets $\varepsilon_k \propto S_k$, the in-domain advantage is a strictly increasing function of ε_k.

Proof. Under the two hypotheses, the release is $\mathcal{N}(\mu_0, \sigma_{\mathrm{DP}}^2 I)$ and $\mathcal{N}(\mu_1, \sigma_{\mathrm{DP}}^2 I)$ with $\mu_1 - \mu_0 = \Delta$. By Neyman–Pearson the optimal test is the likelihood ratio, and $\max(\mathrm{TPR} - \mathrm{FPR})$ over all tests equals the total variation distance between the two laws. Projecting onto $\Delta/\|\Delta\|_2$ reduces this to two univariate Gaussian of common variance separated by $\|\Delta\|_2$, the orthogonal coordinates being identically distributed under both hypotheses and hence uninformative. For that pair $\mathrm{TV} = 2\Phi_{\mathcal{N}}(\|\Delta\|_2/2\sigma_{\mathrm{DP}}) - 1$. □

Lem. 2 closes the gap between $\mathrm{DP}_{\mathrm{pool}}$ and the attack literature. The mean shift our definitions measure is not merely *analogous* to what a membership adversary exploits. It is the sufficient statistic of the optimal such adversary, so the three exposures of Sec. A.2 are one hypothesis test at three granularities and bounding ε_k bounds the power of every distinguisher. The in-domain case differs from the classical one only in its neighborhood: replacing X_k^{sel} by all of $\mathcal{D}_k$ changes many records at once, so it is a group-level statement about an institution's distributional identity rather than about any single patient — which is the unit we set out to protect.

A.2 Three Levels of Privacy Exposure

The heterogeneous multi-institutional setting admits at least three natural notions of what it means for a pooling strategy to be *privacy-preserving*, depending on whether we protect individual patients, the distributional identity of an institution, or the very fact of a hospital's participation. We formalize all three, as each interacts differently with the design of our method.

Sample-level privacy. At the finest granularity we protect individual patient records. Two datasets are *sample-level neighbours* if they differ by exactly one record: $\mathcal{D}' = \mathcal{D} \setminus \{x_i\}$ for some i.

Definition 3 (**Sample-level sensitivity**). *For a selected subset $X_k^{\mathrm{sel}} \subseteq \mathcal{D}_k$ of size n, the sample-level sensitivity with respect to the removal of a single record $x_i \in X_k^{\mathrm{sel}}$ is*

$$\Delta_{\mathrm{samp}}(x_i) \;=\; \big\|\mu(X_k^{\mathrm{sel}}) - \mu(X_k^{\mathrm{sel}} \setminus \{x_i\})\big\|_2 \;=\; \frac{\big\|x_i - \mu(X_k^{\mathrm{sel}})\big\|_2}{n-1}.$$

Plugging $S = \Delta_{\mathrm{samp}}(x_i)$ into $\varepsilon \;=\; \frac{S\sqrt{2\ln(1.25/\delta)}}{\sigma_{\mathrm{DP}}}$ yields the per-record privacy cost. Geometrically, a sample that lies far from the current subset mean induces a large shift in the mean when removed, and is therefore individually privacy-sensitive. In other words, outlier records (patients with unusual feature configurations) are the most dangerous to include under sample-level DP_{pool} (Ex. 3). This is exactly the intuition formalized in standard analyses of the Gaussian mechanism Dwork (2025).

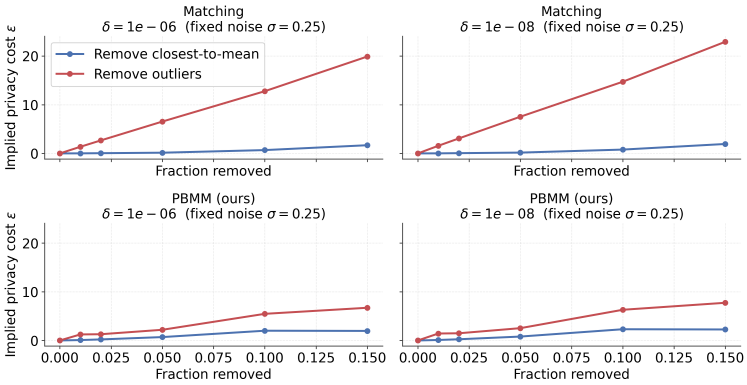


Figure 7: **Sample-level privacy analysis.** Per-sample implied privacy cost ε as a function of the fraction of samples removed, under two removal strategies: *remove closest-to-mean* (samples with smallest $\|x - \mu_k^{\text{all}}\|$) and *remove outliers* (samples farthest from μ_k^{all}). According to Definition 3, the sample-level sensitivity $\Delta_{\text{samp}}(x_i) = \|x_i - \mu(X_k^{\text{sel}})\|_2 / (n-1)$ is governed by how far a sample lies from the current subset mean: samples close to the mean induce a negligible shift upon removal and thus incur low privacy cost, whereas tail samples cause a large mean displacement and are individually privacy-dangerous. Consequently, a pooling strategy that preferentially retains near-mean samples-as PBMM does via its privacy-tempered margin-achieves substantially lower implied ε than one that admits outliers ($\delta = 10^{-6}, 10^{-8}$).

Remark. Sample-level sensitivity scales as $1/n$: larger pooled subsets are more robust to the removal of any one record, as expected from classical concentration and stability arguments in empirical process theory (Vershynin, 2018). This scaling does not, however, neutralize the risk from highly anomalous records. It merely amortizes their influence across the pool. In particular, a single point far in the tail of the distribution can still dominate $\Delta_{\text{samp}}(x_i)$. This is exactly the observation in Fig. 7

In-domain (hospital-population) privacy. Sample-level privacy protects individual patients, but does not prevent leakage about the *distributional identity* of the contributing institution. Even if no single record can be recovered from the released mean, a highly skewed selection can reveal that a hospital's underlying population harbours an unusual subgroup, rare condition prevalence, or systematic demographic imbalance. To capture this, we study a domain-level analogue of sensitivity. Two datasets are *in-domain neighbours* for institution k if they differ by replacing the selected subset X_k^{sel} with the hospital's full domain $\mathcal{D}_k$. Intuitively, we ask *how much the released statistic would change if the institution had contributed all of its data instead of the subset chosen by pooling*.

Definition 4 (In-domain sensitivity). *For a selected subset $X_k^{\text{sel}} \subseteq \mathcal{D}_k$, the in-domain sensitivity of institution k is $S_k = \left\|\mu(X_k^{\text{sel}}) - \mu_k^{\text{all}}\right\|_2$, where $\mu_k^{\text{all}} = n_k^{-1} \sum_{x \in \mathcal{D}_k} x$ is the true population mean of institution k. The corresponding (ε_k, δ)-DP_{pool} cost under the Gaussian mechanism is given by $\varepsilon_k = \frac{S_k \sqrt{2\ln(1.25/\delta)}}{\sigma_{\text{DP}}}$.*

Here S_k measures how atypical the selected subset is relative to the institution's own population. A small S_k indicates that the selection is statistically representative of the hospital; the released mean reveals little beyond what would already be inferred from an i.i.d. sample. A large S_k indicates that the selection is drawn from a thin shell of the hospital's distribution, and thus leaks structural information about its population.

Proposition 1 (Privacy ordering of radial bands). *Let μ_k^{all} denote the population mean of institution k and $R_k = \max_{x \in \mathcal{D}_k} \|x - \mu_k^{\text{all}}\|_2$ its maximal radius. Partition $\mathcal{D}_k$ into concentric shells around μ_k^{all} with inner radius $R_k/3$ and outer radius R_k:*

$$\mathcal{B}_{\text{in}} = \left\{x : \|x - \mu_k^{\text{all}}\|_2 \le R_k/3\right\}, \quad \mathcal{B}_{\text{mid}} = \left\{x : R_k/3 \le \|x - \mu_k^{\text{all}}\|_2 \le 2R_k/3\right\},$$

$$\mathcal{B}_{\text{out}} = \left\{x : 2R_k/3 \le \|x - \mu_k^{\text{all}}\|_2 \le R_k\right\}.$$

Then, for any selection X_k^{sel} drawn entirely from a single band whose displacements are radially aligned in the sense of Assumption 1, i.e. $x - \mu_k^{\text{all}} = r\, u_y$ with $r \ge 0$ (as the matching gate favours, since it rewards motion along u_y alone), we have

$$S_k(\mathcal{B}_{\text{in}}) \le S_k(\mathcal{B}_{\text{mid}}) \le S_k(\mathcal{B}_{\text{out}}).$$

Proof. Let $\mu := \mu_k^{\text{all}}$ and define the radius function $r(x) := \|x - \mu\|_2$. By construction, all points in $\mathcal{D}_k$ lie in the closed ball $\overline{B}(\mu, R_k)$. We use the fact that the mean of a finite set lies in its convex hull,

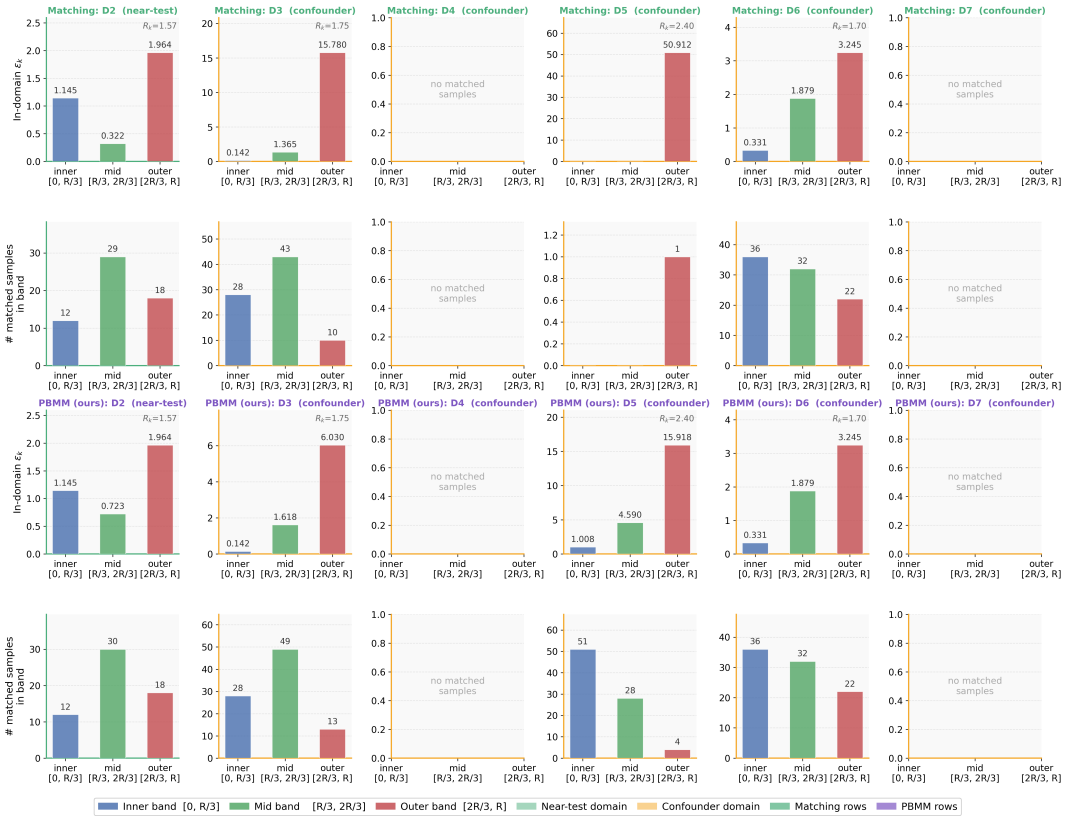


Figure 8: **In-domain privacy cost ε_k and sample counts across radial bands.** Each column corresponds to a non-anchor institution D_k; rows alternate between implied privacy cost ε_k (top) and matched sample count (bottom) for Matching (rows 1–2) and PBMM (rows 3–4). Bars are coloured by radial band: inner $[0, R_k/3]$ (blue), mid $[R_k/3, 2R_k/3]$ (green), and outer $[2R_k/3, R_k]$ (red), where $R_k = \max_{x \in \mathcal{D}_k} \|x - \mu_k^{\text{all}}\|_2$. For Matching, confounder institutions (D_3–D_7) concentrate their selected samples in the outer band, incurring the highest ε_k (e.g., D_5: outer $\varepsilon_k = 50.9$) while the inner band remains empty-a direct empirical consequence of Theorem 1 (matching is contravariant under the privacy metric). PBMM's privacy-tempered margin redistributes admissions toward the inner and mid bands: compare D_5 under PBMM, where 51 inner-band and 28 mid-band samples are admitted against only 4 outer-band samples, yielding a substantially lower peak $\varepsilon_k = 15.9$ versus 50.9 for Matching. Institutions with no matched samples (D_4, D_7) are hard-excluded by Stage 1 (Section 3.2.1), confirming that the matched core acts as a necessary gateway for privacy-tempered expansion.

and that convex combinations are contractive with respect to convex sets in Euclidean space (see, e.g., Vershynin (2018, Chapter 1)).

Inner band. For any $x \in \mathcal{B}_{\text{in}}$, we have $r(x) \le R_k/3$, so $\mathcal{B}_{\text{in}} \subset \overline{B}(\mu, R_k/3)$. Since closed Euclidean balls are convex, the convex hull of $\mathcal{B}_{\text{in}}$ is also contained in $\overline{B}(\mu, R_k/3)$. Therefore, for any finite subset $X_k^{\text{sel}} \subset \mathcal{B}_{\text{in}}$, $\mu(X_k^{\text{sel}}) \in \text{conv}(X_k^{\text{sel}}) \subset \overline{B}(\mu, R_k/3)$, which implies $\|\mu(X_k^{\text{sel}}) - \mu\|_2 \le R_k/3$.

Mid and outer bands. Under radial alignment every admitted displacement is $r\,u_y$ with r in the band's radius interval, so $\mu(X_k^{\text{sel}}) - \mu_k^{\text{all}} = \bar{r}\,u_y$ with $\bar{r}$ the mean radius, giving $S_k(X_k^{\text{sel}}) = \bar{r} \in [\rho_{b-1}, \rho_b]$. Since $[0, R_k/3]$, $[R_k/3, 2R_k/3]$ and $[2R_k/3, R_k]$ are ordered, so are the corresponding sensitivities.

Without radial alignment. The upper bounds survive unconditionally: every band lies in the closed ball of its outer radius, which is convex, so $S_k(X) \le \rho_b$ for any selection from band b, and the bound is attained by a singleton at radius ρ_b. The worst-case cost is therefore ordered across bands with no assumption at all, which is what Stage 3's admission order requires. □

Proposition 1 states that, holding the number of contributed samples fixed, an institution minimizes its in-domain privacy exposure by selecting samples close to its own mean μ_k^{all}. In contrast, matching-based pooling (Roy et al., 2026) admits samples that are close to the *anchor's* class centroids. For a shifted institution, these are precisely the samples farthest from μ_k^{all}. This *tail concentration* is the fundamental privacy failure mode of matching that our method aims to correct.

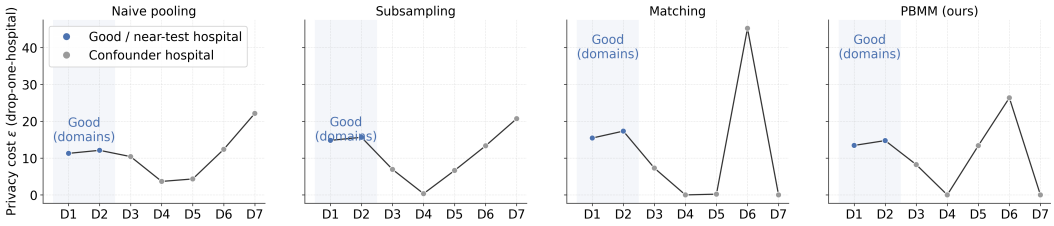


Figure 9: **Hospital-level privacy analysis.** Drop-one-hospital implied privacy cost ε for each method, computed via a *remove-and-replace* protocol: institution D_k is first removed from the pool ($\mathcal{P}_{-k}$), the pooled mean $\mu(\mathcal{P}_{-k})$ is recorded, institution k's contribution is then re-admitted to restore $\mathcal{P}$, and the hospital-level sensitivity $\Delta_{\text{hosp}}(k) = \frac{n_k^{\text{sel}}}{N}\|\mu(X_k^{\text{sel}}) - \mu(\mathcal{P}_{-k})\|_2$ is computed as the shift induced by reinserting D_k (Definition 5). Good/near-test hospitals (shaded region) contribute samples close to the anchor distribution and thus cause only a modest mean displacement upon removal or reinsertion, whereas confounder hospitals-whose selected subsets are distributionally atypical-induce large shifts and high ε. PBMM's radial admission order (inner $\rightarrow$ mid $\rightarrow$ outer band) ensures that each institution's contribution stays close to its own population mean μ_k^{all}, which simultaneously keeps $\Delta_{\text{hosp}}(k)$ small and limits the exposure of an institution's participation to an external adversary.

Note. The directional consistency assumption ensures that all admitted samples satisfy $\langle x - \mu_k^{\text{all}}, u_y \rangle \geq 0$, confining X_k^{sel} to a half-space. The convex hull of any finite subset of this half-space intersected with $\mathcal{B}_{\text{out}}$ therefore cannot contain μ_k^{all}, precluding cancellation and ensuring $\|\mu(X_k^{\text{sel}}) - \mu_k^{\text{all}}\|_2 \geq R_k/3$.

Fig. 8 empirically corroborates Definition 4: across all non-anchor institutions, the per-band privacy cost ε_k is strictly ordered inner $\leq$ mid $\leq$ outer (Proposition 1), and PBMM's radial admission order provably shifts the sample-count mass toward the inner band-reducing peak in-domain sensitivity relative to Matching (Roy et al., 2026) while retaining or increasing the total number of admitted samples. [4]

Hospital-level (participation) privacy. Another aspect of DP_{pool} involves the most stringent notion protecting not individual samples or population statistics, but the very fact of an institution's participation in the pooled dataset. Two datasets are *hospital-level neighbors* if they differ by the removal of one institution's entire contribution: $\mathcal{D}' = \mathcal{D} \setminus \mathcal{D}_k$.

Definition 5 (**Hospital-level sensitivity**). *Let $\mathcal{P}$ denote the pooled dataset and $\mathcal{P}_{-k} = \mathcal{P} \setminus X_k^{\text{sel}}$ the pool with institution k removed. The hospital-level sensitivity is*

$$\Delta_{\text{hosp}}(k) \;=\; \|\mu(\mathcal{P}) - \mu(\mathcal{P}_{-k})\|_2 \;=\; \frac{n_k^{\text{sel}}}{N}\,\|\mu(X_k^{\text{sel}}) - \mu(\mathcal{P}_{-k})\|_2,$$

where $n_k^{\text{sel}} = |X_k^{\text{sel}}|$ and $N = |\mathcal{P}|$ are the institution's contribution and total pool size, respectively.

Hospital-level sensitivity is driven by two compounding factors: (i) the *relative size* of the institution's contribution and (ii) the *distance* between its selected mean and the mean of the remaining pool. An institution that contributes many atypical samples is doubly exposed: its removal causes a large shift in the pooled mean simply by volume, and that shift is further amplified by the distributional mismatch (Roy et al., 2026; Chytas et al., 2024).

Fig. 9 empirically validates Definition 5: across all four pooling strategies, the drop-one-hospital privacy cost ε is consistently lower for good/near-test institutions than for confounder institutions, and PBMM's monotone radial expansion yields the smallest sensitivity peaks, confirming that controlling in-domain centrality (Sec. 3.2.2) provides a principled handle on participation-level privacy. The utility consequence of this privacy ordering is made explicit in Fig. 10: institutions that carry high hospital-level sensitivity $\Delta_{\text{hosp}}(k)$ (Fig. 9) are precisely those whose removal produces

[4] In the heterogeneous and homogeneous real-data settings (Sec. B.3, B.2), privacy costs are computed in $\mathbb{R}^{d_\Phi}$ where $d_\Phi = 768$ (CLIP feature space) and $d_\Phi = 741$ (ABIDE correlation matrix space) respectively. Under a fixed $\sigma_{\text{DP}} = 0.25$ calibrated for the 2D synthetic setting, the implied ε values in Figs. 15 and 16 appear large, a known consequence of applying a fixed noise scale across spaces of differing dimensionality Dwork & Roth (2014), where ℓ_2 sensitivity naturally grows with d_Φ. A dimension-aware rescaling $\hat{\varepsilon}_k = \varepsilon_k/\sqrt{d_\Phi}$ would bring values into a comparable range, but we report raw scores throughout since all relative orderings (inner $\leq$ mid $\leq$ outer) and cross-method comparisons are invariant to this uniform rescaling.

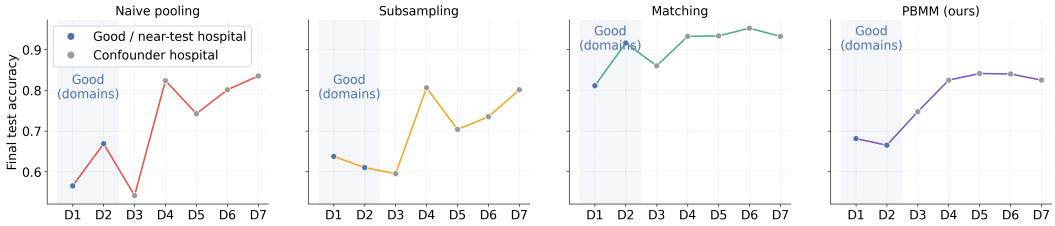


Figure 10: **Leave-one-domain-out performance analysis.** Final test accuracy after removing each institution D_k in turn, retraining on the remaining pool, and evaluating on the fixed test distribution $P^\star$. Good/near-test institutions (shaded region, D_1–D_2) are the anchor-aligned domains whose supervision signal is consistent with $P^\star$; confounder institutions (D_3–D_7) apply conflicting label permutations. Crucially, the performance curve is the *utility dual* of the privacy cost curve in Fig. 9: a domain whose removal causes a *large drop* in accuracy is precisely a domain that was contributing distributionally central, label-consistent samples-the same samples that, by Definition 5 and Proposition 1, carry the *lowest* hospital-level sensitivity $\Delta_{\text{hosp}}(k)$. Conversely, confounder domains whose removal causes accuracy to *rise* (visible in Naïve pooling and Subsampling) are those whose selected subsets lie far from the anchor distribution, inducing large $\Delta_{\text{hosp}}(k)$ spikes in Fig. 9. PBMM's radial admission order keeps each institution's contribution close to μ_k^{all}, simultaneously suppressing $\Delta_{\text{hosp}}(k)$ *and* stabilizing the accuracy curve: removing any single institution causes only a modest performance change, indicating that no institution is either indispensable (privacy-dangerous) or harmful (utility-degrading).

the largest accuracy swings, either a sharp drop when a label-consistent, central-mean institution is removed, or a counter-intuitive *gain* when a confounder institution is removed, confirming that $\Delta_{\text{hosp}}(k)$ serves as a reliable, privacy-grounded predictor of an institution's net utility contribution to the pool. More precisely, since the downstream MLP's decision boundary is anchored to the empirical class centroids of the pooled training set, removing institution D_k displaces those centroids by a quantity proportional to $\Delta_{\text{hosp}}(k) = \frac{n_k^{\text{sel}}}{N} \|\mu(X_k^{\text{sel}}) - \mu(\mathcal{P}_{-k})\|_2$; this centroid displacement directly rotates the learned decision boundary away from (or, for confounders, towards) the test geometry of $P^\star$, making $\Delta_{\text{hosp}}(k)$ not merely a statistical proxy for utility impact but the *direct geometric cause* of the accuracy change observed in Fig. 10 and simultaneously the quantity our privacy budget controls.

Remark. Hospital-level privacy subsumes in-domain privacy in the following sense. If a pooling strategy keeps S_k small for each institution (selecting samples close to μ_k^{all}), and if the pool is dominated by institutions whose μ_k^{all} are well-aligned with the target distribution (as ensured by matching (Roy et al., 2026)), then $\mu(X_k^{\text{sel}})$ lies close both to μ_k^{all} and to $\mu(\mathcal{P}_{-k})$. Under mild regularity conditions on the mixture, this implies $\Delta_{\text{hosp}}(k)$ is also small. In other words, controlling in-domain sensitivity provides a principled handle on participation-level DP when matching has already enforced distributional alignment.

A.3 Related Work on Privacy-Utility Trade-offs and Training-based privacy preserving methods

Harmonization and Invariance. In standard multi-site analyses, it is common practice to control for nuisance variables such as scanner type, race, or gender to ensure that genuine features drive predictions rather than site-specific artifacts. In the context of deep learning, this harmonization challenge is frequently addressed through invariant representation learning Moyer et al. (2020; 2018). These approaches attempt to map inputs to a latent space that is explicitly invariant to certain covariates or sensitive attributes (Chytas et al., 2024; Lokhande et al., 2022; 2020; Yang et al., 2023; Stojanov et al., 2021). *However, relying solely on latent-space adjustments is often insufficient as all covariates/variables cannot be addressed practically Luo & Ren (2024).* This highlights the need to resolve this problem in the data space where the problem originates.

Advanced Data Pooling. Recent work has moved beyond naive aggregation toward filtering-based strategies that remove confounding samples likely to distort the target distribution (Roy et al., 2026). In parallel, relaxing the i.i.d. assumption and modeling exchangeability across sites improves harmonization and mitigates distribution shifts as more sources are added (Roy et al., 2025; Shen et al., 2024). These approaches emphasize that effective pooling depends on *which* samples are selected, not just *how many*. However, they overlook a critical dimension: the explicit *privacy cost* of including samples under a formal budget. Ignoring this is not merely a statistical issue as model

reliability is highly sensitive to training data distributions, with harder tasks exposing high-variance, incoherent failure modes when data is misaligned Hägele et al. (2026).

Privacy-Enhancing Techniques. Numerous *privacy-enhancing technologies* (PETs) have been proposed for collaborative data use without centralizing raw records. *Federated learning* McMahan et al. (2017) allows institutions to share only model updates rather than data, but remains susceptible to gradient-inversion and membership-inference attacks Hatamizadeh et al. (2022). *Secure multi-party computation* (MPC) enables joint computation over private inputs with provable security, but communication overheads are often prohibitive for the feature-matching operations required in data pooling. *Homomorphic encryption* permits arithmetic on encrypted data but similarly struggles with the dimensionalities typical of medical imaging. Crucially, these approaches do not address *sample-selection*: which data points to include in a pooled training set under distributional heterogeneity to address a finite privacy budget. This motivates a privacy notion for the pooling stage itself. We define $\mathrm{DP}_{\mathrm{pool}}$ (Def. 2): the (ε, δ) guarantee of Def. 1 evaluated at the neighboring pair the pooling protocol actually produces, rather than at the supremum over all conceivable pairs. Lemma 2 shows this is the same inequality and controls the same attack; the quantifier is the only difference, and Sec. A.1 gives the design change that restores it.

Privacy-Utility tradeoff. The tension between privacy and utility under differential privacy has been studied across a range of theoretical and applied settings. Hamm (2015) introduces minimax filters that learn a privacy-utility optimal transformation of continuous high-dimensional data by formulating the problem as a three-party game between a data contributor, a cooperative aggregator, and an adversary. *However, their approach operates entirely in a learned representation space and does not address multi-site data pooling, sample selection under domain shift, or per-institution budgeting*. Zhong & Bu (2022) provide a rigorous information-theoretic characterization of the privacy-utility trade-off (PUT) under several notions of DP, including approximate-DP and maximal leakage, which contextualizes the Lagrangian framing of our composite score m_{priv} (Eq. 4). *Unlike PBMM, their analysis concerns the design of randomized mechanisms over fixed datasets rather than geometry-aware, budget-enforced sample selection across heterogeneous institutions*. Padakandla et al. (2019) precisely characterize the utility-privacy trade-off for database sanitizing mechanisms that *add calibrated noise to perturb released statistics*, using tools from Ehrhart theory in the asymptotic regime of large databases. However, as noted in Sec. 1, noise-based perturbations distort the data manifold, degrading calibration and inducing covariate shifts that harm generalization in downstream models Müller et al. (2019); Rosenfeld et al. (2020). PBMM takes the conceptually opposite approach: rather than perturbing any sample, it *selects* which unmodified samples to admit based on in-domain centrality, controlling sensitivity geometrically without corrupting the supervision signal-a distinction that is especially critical in data-scarce medical imaging where label fidelity directly determines clinical utility. Bailie et al. (2024) examine conundrums that arise when differential privacy is combined with *statistical invariants*-pre-specified aggregate quantities (e.g., census marginals) that must be released without noise. This notion of invariance is structurally distinct from the *distributional harmonization* objective in PBMM, which seeks geometric alignment of pooled samples with a target domain under heterogeneous shift rather than preservation of fixed summary statistics. Critically, they identify the DP-invariance tension as an open theoretical conundrum and do not provide an algorithm to resolve it; PBMM operationally resolves the analogous utility-privacy tension through the minimax matching rule and monotone radial expansion with a provable budget guarantee. On the applied side, Galada et al. (2025) propose PRISM, a privacy-preserving inter-site MRI harmonization framework based on disentangled representation learning that harmonizes without raw data exchange. *Unlike PBMM, PRISM does not enforce an explicit per-institution (ε, δ)-DP budget, does not address the three-way trade-off between harmonization, sample utilization, and privacy, and operates entirely in a learned latent space rather than performing sample selection in the data space*. Nanayakkara et al. (2022) offer tools for reasoning about and visualizing the probabilistic privacy-utility trade-off to guide a curator's *choice* of ε before deployment. PBMM, by contrast, takes $\varepsilon_k^{\max}$ as a hard per-institution constraint already fixed by regulatory or institutional agreement, and makes *geometric sample admission decisions within it* at pooling time-*a fundamentally different problem that their visualization framework has no mechanism to address, particularly under heterogeneous domain shift where the cost of each admission depends on its radial distance from the institution's own population mean.* kim et al. (2025) propose BboxHarmony, a privacy-preserving MRI harmonization framework that transforms a *single* institution's target images toward an unknown source domain style via Bayesian optimization over a disentanglement-based generator, without data sharing or model parameter access. Unlike PBMM, it operates on one institution in isolation without modeling cross-

institutional domain shift, enforces no formal (ε, δ)-DP budget, and modifies samples through style transfer rather than selecting unmodified samples based on in-domain representativeness-meaning it addresses the *output side* of privacy (what to release) rather than the *input side* (which training samples to pool under a differential privacy guarantee).

Training-time DP techniques. A parallel line of work incorporates differential privacy directly into the training procedure, most notably through DP-SGD Abadi et al. (2016), which clips and noises per-example gradients to bound the influence of any single training point on model parameters, and PATE Papernot et al. (2016), which aggregates noisy votes from teacher models trained on disjoint private partitions. While these methods provide formal (ε, δ)-DP guarantees on the *model*, they are structurally inapplicable to the problem PBMM addresses for two reasons. First, they assume a fixed, pre-collected training set and inject noise into the *learning algorithm*; they offer no mechanism to decide *which samples to include* in that set when the candidate pool is heterogeneous and confounded, which is precisely the pooling problem. Second, noise injection during training distorts the supervision signal and degrades calibration Müller et al. (2019), which is especially damaging in data-scarce medical settings where label fidelity is non-negotiable Rosenfeld et al. (2020). PBMM instead enforces privacy at the *data selection stage* itself: it controls which unmodified samples enter the pool so that the in-domain sensitivity S_k stays within a per-institution budget, leaving the downstream training algorithm entirely unconstrained and model-agnostic. *This separation is deliberate-privacy is budgeted where the problem originates, in the data space, rather than compensated for after the fact during optimization*.

Instance-specific privacy accounting. $\mathrm{DP}_{\mathrm{pool}}$ (Def. 2) belongs to a line of work that reports realized rather than supremal privacy loss. Wang (2019) formalizes per-instance differential privacy, the loss incurred by a specific dataset under a fixed mechanism, and Nissim et al. (2007) introduce smooth sensitivity, which calibrates noise to a data-dependent envelope. We differ in what the quantity is used *for*: both are accounting or calibration tools applied to a given dataset, whereas $\mathrm{DP}_{\mathrm{pool}}$ is a constraint on which samples are admitted, so tightening it changes the selection rather than the noise. The in-domain neighborhood (Def. 4) is also specific to pooling: the protected unit is an institution's distributional identity, which has no analogue in the single-dataset setting those works consider.

B Experimental Setups

B.1 Synthetic Gaussian Data Addition

Feature geometry. We construct a controlled multi-domain benchmark in $\mathbb{R}^2$ to isolate the effects of domain heterogeneity on pooling quality and privacy exposure. Class blob centres are placed on a circle of radius $r = 2.2$ in the feature plane:

$$\mathbf{t}_c = \big(r\cos(2\pi c/C),\ r\sin(2\pi c/C)\big), \qquad c = 0, \ldots, C-1, \tag{10}$$

yielding two diametrically opposed blobs for $C = 2$ (left/right), an equilateral triangle for $C = 3$, and so on. Within-blob spread is isotropic Gaussian with standard deviation $\sigma_{\text{blob}} = 0.60$.

Domain structure. We generate $K = 8$ domains in total: one anchor domain $\mathcal{D}_a$, two additional *near-test* domains ($\mathcal{D}_2, \mathcal{D}_3$), and five *confounder* domains ($\mathcal{D}_4, \ldots, \mathcal{D}_8$). Near-test domains share the identity label permutation with the target (blob at position $\mathbf{t}_c$ carries label c) with 2% random label noise and $n_{\text{good}} = 60$ samples per class. Confounder domains apply distinct permutations of $[C]$ to the same blob positions, so the same feature region receives conflicting supervision signals across institutions. To simulate realistic size heterogeneity, confounder domains are substantially larger: $n_{\text{conf}} = \lfloor 300 \times \eta \rfloor$ samples per class, where $\eta = 1.5$ is the *heterogeneity parameter*. At $\eta = 1.5$ (which is a sufficiently strong form of domain heterogeneity as seen in Fig. 13) each confounder contributes 90 samples per class, more than five times the near-test domains, sufficient to overwhelm naïve pooling. Label noise in confounder domains is set to 3%. *Confounded domains are deliberately designed to have more number of samples (90 samples) than the good domains (60 samples). This mimics the real world scenarios: well annotated error free data curation requires time and money which limits the size whereas cheaper faster curation leads to error prone datasets.*

Test distribution. The test set consists of 500 samples per class drawn from the identity-permutation blobs with no label noise, giving $500 \times C$ test samples in total. The test distribution $P^\star$ is never accessed during pooling; it is used solely for evaluating downstream classifier performance.

Label permutation. The confounders use cyclic shifts, reversals, and random shuffles of $[C]$ as their labeling conventions; no two confounders agree on the same wrong answer (see Fig. 13). This design stress-tests all pooling strategies simultaneously: naïve pooling sees contradictory supervision for every feature region; subsampling draws randomly from both correct and incorrect labels; matching, anchored on $\mathcal{D}_a$, initializes class centroids at the correct blob positions and filters confounder samples whose labels disagree with this geometry. This adds to the already complex heterogeneous domain shifted setup.

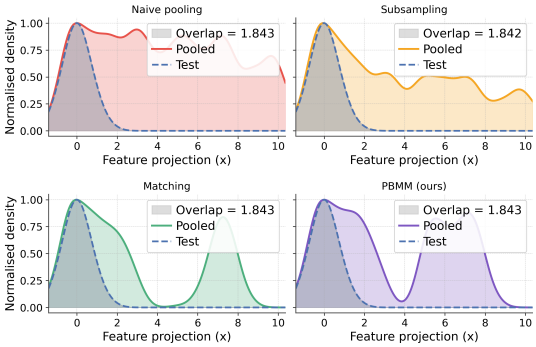


Figure 11: **Pooled distribution vs. Test distribution.** Filtering-based pooling techniques (matching and PBMM) outperform traditional techniques.

Classifier and training protocol. We train a two-hidden-layer MLP (widths 128 and 64, ReLU activations, Adam optimiser, learning rate 5×10^{-4}, ℓ_2 regularisation 10^{-3}) for 80 epochs on each pooled dataset, measuring test accuracy at every epoch. This aligns to non-degrading downstream performance observation in (Roy et al., 2026; 2025; Shen et al., 2024) Features are standardised to zero mean and unit variance before training. All experiments use a fixed random seed for reproducibility.

Privacy evaluation. In-domain privacy cost is measured via Definition 4 with Gaussian mechanism parameters $\sigma_{\mathrm{DP}} = 0.25$ and $\delta = 10^{-6}$. For each non-anchor domain k and each radial band $b \in \{\mathrm{in}, \mathrm{mid}, \mathrm{out}\}$, we report

$$\varepsilon_k^{(b)} = \frac{S_k^{(b)} \sqrt{2 \ln(1.25/\delta)}}{\sigma_{\mathrm{DP}}}, \quad S_k^{(b)} = \left\| \mu\left(X_k^{\mathrm{sel}} \cap B_b\right) - \mu_k^{\mathrm{all}} \right\|_2 \tag{11}$$

where bands $\mathcal{B}_b$ are defined relative to μ_k^{all} (Proposition 1), consistent with the S_k definition throughout. This band-level decomposition allows us to verify empirically that PBMM's monotone radial admission order (Eq. 6) translates into a monotone privacy cost profile across bands.

Baselines. We compare four pooling strategies: (i) **Naive pooling** $\mathcal{P}_{\mathrm{naive}} = \bigcup_k \mathcal{D}_k$; (ii) **Uniform subsampling** with $n_{\mathrm{sub}} = 80$ samples per class per institution; (iii) **Matching** (Roy et al., 2026) anchored at $\mathcal{D}_1$; and (iv) **PBMM (ours)**. All methods use the same classifier and training protocol. The anchor $\mathcal{D}_1$ is a near-test domain, so matching operates under favourable conditions; this choice benefits the matching baseline, making any privacy improvements of PBMM conservative estimates.

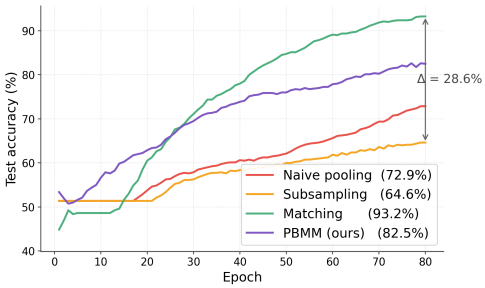


Figure 12: **Downstream performance on pooled set.** Filtering-based pooling techniques (matching and PBMM) outperform traditional techniques in downstream classification (Fig. 11).

Empirical results. It can be seen that filtering-based techniques achieve better data harmonization (Fig. 11) which translates to downstream performance gains (Fig. 12).

Please see Sec. B.4 for additional results.

B.2 Homogeneous Data Addition

Dataset. We use the Autism Brain Imaging Data Exchange (ABIDE) Nielsen et al. (2013) cohort, a publicly available multi-site neuroimaging dataset comprising resting-state fMRI scans from subjects diagnosed with Autism Spectrum Disorder (ASD) and matched neurotypical controls. Each subject's time-series is parcellated into 39 regions using the MSDL atlas. Per-subject connectomes are estimated via Oracle Approximating Shrinkage (OAS) covariance and normalized to correlation matrices, then vectorized by extracting the upper triangle, yielding a feature vector of dimension 741 (see Fig. 14).

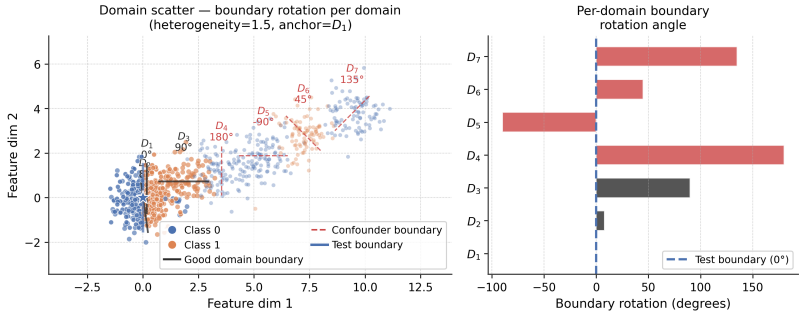


Figure 13: **Synthetic Gaussian experimental setup.** The experimental setting clearly indicates few good domains (domains that are closely aligned with the test domain) and higher number of confounder domains. The decision boudaries of the confounder domains are highly dissimilar to the test and good domains which creates additional difficulty in data harmonization while pooling.

Protocol. Following the ordered-domain homogeneous protocol, $\mathcal{D}_1$ is the fully held-out target site (LEUVEN_1 by default as seen in Fig. 14, selected for having a balanced class distribution and sufficient size). The remaining sites are treated as source domains and ordered by decreasing geometry compatibility score to $\mathcal{D}_1$, so $\mathcal{D}_2$ is the site most similar to the target and later domains are progressively farther away. Training proceeds in a continual fashion: domains are added from the farthest (most dissimilar) source back toward $\mathcal{D}_2$, so the closest-to-target domain is always added last. This ordering mirrors the practical scenario of starting from a weakly related pool and progressively adding more relevant data.

Classifier and training protocol. A compact multilayer perceptron is trained from scratch on the vectorized 39×39 connectome correlation matrices (1521-dimensional input). The architecture comprises a hidden layer of width 128 with LayerNorm, ReLU, and dropout ($p = 0.1$), followed by a linear projection head that produces a 64-dimensional embedding (ReLU + dropout, $p = 0.1$), and a two-class linear classification head. We run training for 20 epochs per curriculum stage with a batch size of 16, using the AdamW optimizer (learning rate 10^{-3}, weight decay 10^{-4}). At each domain-addition stage, evaluated under the `d2_closest` and `d2_farthest` setups targeting the LEUVEN_2 site, PBMM and Matching compute selection in the *embedding space* of the current MLP encoder (see Algorithm 1). Privacy costs are computed as in Sec. B.1, with privacy parameters $\sigma_{\text{DP}} = 0.25$ and $\delta = 10^{-6}$.

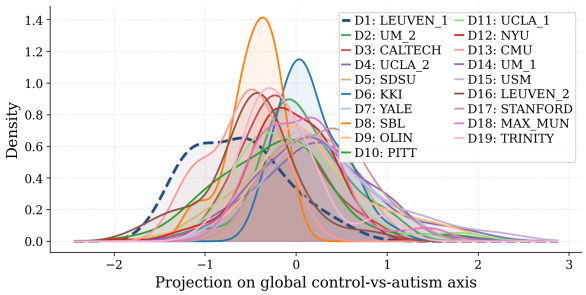


Figure 14: **Homogeneous ABIDE setup.** D_1 (LEUVEN_1) is the fully held-out target site due to its balanced class distribution and sufficient size. The remaining sites are treated as source domains and ordered by decreasing geometry-compatibility score with respect to D_1, so D_2 is the closest source and later domains are progressively farther away. Training proceeds in the ordered-domain protocol by adding sources from the farthest back toward D_2.

Baselines. Baselines are same as Sec. B.1

Experimental Results. We see the expected trend in Table 1 where filtering-based techniques, i.e., Matching and PBMM, outperforms Naive Pooling and Subsampling with respect to DA score. Furthermore, PBMM also achieves comparable performance to Matching, thus maintaining utility while ensuring better data harmonization (DA score).

B.3 Heterogeneous Data Addition

Dataset. We pool six heterogeneous medical imaging datasets (BMAD Bao et al. (2024)) spanning distinct modalities: brain MRI anomaly detection, liver CT lesion segmentation, retinal fundus imaging (two cohorts: RESC and OCT2017), chest X-ray pathology detection, and histopathology slide analysis. Collectively, these datasets cover a wide range of tissue types, imaging physics, and acquisition devices, making direct geometric comparison in image space meaningless.

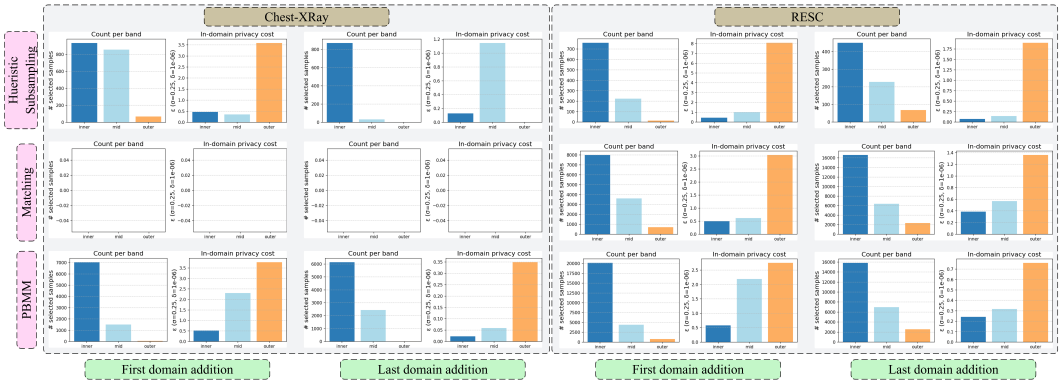


Figure 15: **Privacy-band analysis in the heterogeneous zero-shot setting.** We compare Matching and PBMM across the inner, mid, and outer privacy bands induced by the radial ordering around the matched core. Across datasets, PBMM shifts admissions toward the cheaper inner and mid bands while reducing reliance on outer-band samples, thereby lowering overall privacy exposure without sacrificing harmonization performance. Matching, being purely filtering-based, tends to over-filter and under-utilize available samples, which is especially visible on datasets such as Chest-XRay.

Model and training. A frozen CLIP ViT-L/14 backbone Radford et al. (2021) with lightweight adapter modules serves as encoder Φ. Only adapter weights are updated during training. Anomaly detection and segmentation heads operate on patch-token features from multiple intermediate layers. PBMM selection uses Algorithm 1 at each domain-addition stage: core-matching and privacy-tempered margins are computed in the 768-dimensional CLIP space, with DP parameters $\sigma_{\mathrm{DP}} = 0.25$, $\delta = 10^{-6}$ (consistent with Setups 1–2). The setup is same as Huang et al. (2024); Roy et al. (2026).

Baselines. We utilize matching and naive pooling from Roy et al. (2026) like the other setups. We reinterpret sub-sampling into a heuristic based subsampling technique Navidi et al. (2025). Uniform subsampling (which draws at most $\lfloor n_{\mathrm{sub}}/C \rfloor$ samples per class per institution without regard to distributional geometry) is the classical privacy-agnostic size-balancing heuristic. We replace it with a stronger, geometry-aware variant inspired by the adaptive resampling framework of Navidi et al. (2025), who show that reweighting samples according to their density in a learned latent space substantially improves representation of underrepresented and out-of-distribution cellular states in imbalanced single-cell data. We adapt this principle to our pooling setting: rather than drawing uniformly, we subsample each institution proportionally to inverse local density estimated in feature space, biasing selection toward atypical, informative samples. This is the *direct geometric opposite* of filtering based techniques-where PBMM and matching preferentially admits samples close to an institution's population mean (small S_k), heuristic subsampling deliberately seeks out samples far from the bulk of the distribution, maximizing coverage at the expense of in-domain representativeness. The comparison therefore provides a clean ablation of the privacy-utility axis: both methods go beyond naive uniform draws, but in opposing directions along the radial spectrum formalized by Proposition 1.

Empirical results. It can be seen that PBMM outperforms all baselines in terms of privacy (Fig. 15) while sustaining only an acceptable dip in distributional alignment score (Fig. 4), remaining competitive with existing pooling techniques on downstream accuracy. Notably, heuristic subsampling achieves strong coverage of shifted domains but incurs the highest in-domain sensitivity S_k of any method, confirming that latent-space diversity and in-domain privacy are structurally in tension which is precisely the incompatibility proven in Theorem 1.

B.4 Data-Scarce Sequential Pooling

Datasets. We utilize an in-house dataset as the anchor dataset. It contains 43 subjects. We partition them into three age groups: Younger (< 30 years), Midlife ($30 - 54$ years), and Older (≥ 55). All the patients are healthy control group patients with no neuro-degenerative diseases. For the sequential data addition, we utilize ADNI Petersen et al. (2010). We retain one scan per participant and that produces 1013 unique participants in our cohort. After the age grouping is performed, in-house dataset contains 9 younger, 23 midlife, and 11 older subjects whereas ADNI contains 3 midlife, 1010 older, and no younger participants. So only midlife and older groups contribute to the external

samples. For each participant, we estimate a 39×39 functional-connectivity matrix (MSDL-39 ROI Atlas) and retain its 741 strict upper-triangular coefficients. Features are standardized using the in-house training statistics. Within each fold, ADNI connectomes are harmonized by mapping their element-wise first and second moments to those of the in-house training partition.

Sequential Pooling Protocol. For each random seed in $\{1, 7, 13, 19, 31, 42\}$, we construct nested ADNI subsets containing 15, 30, 45, 60, 75, and 90 participants. Evaluation uses stratified three-fold cross-validation over the in-house dataset, with all preprocessing statistics estimated exclusively from the corresponding in-house dataset's training fold. Reported values are averaged across folds and seeds.

Model and training. We use a two-layer MLP with architecture $741 \to 32 \to 16 \to 3$ and GELU activations, without dropout or LayerNorm. The model is trained for 90 epochs using AdamW with learning rate 2×10^{-3} and weight decay 2×10^{-4}. Training follows the natural pooled distribution without class-balanced sampling.

Baselines. Same as Sec. B.2.

C Theoretical Results, Algorithmic Explanations and Ablations of PBMM

C.1 Proof of Lemma 1

Lemma 1 (Matched-core mean as a proxy under sufficient overlap). *Let $D_k = M_k \sqcup L_k$ with $\alpha_k := |M_k|/|D_k| \in (0, 1]$, $\hat{\mu}_k := \mu(M_k)$, $\mu_k^{\mathrm{all}} := \mu(D_k)$, and $\mu_{L,k} := \mu(L_k)$. Assume institution k is τ_k-overlapping with the anchor-consistent region in the sense that $\|\hat{\mu}_k - \mu_{L,k}\|_2 \le \tau_k$. Then $\|\hat{\mu}_k - \mu_k^{\mathrm{all}}\|_2 \le (1 - \alpha_k)\tau_k$. In particular, if matching returns a substantial slice (α_k not small) and the matched slice lies in a local overlap region of radius τ_k, then $\hat{\mu}_k$ is a good proxy for the full institution mean μ_k^{all}.*

Proof. Since $D_k = M_k \sqcup L_k$, we have the convex decomposition

$$\mu_k^{\mathrm{all}} = \alpha_k \hat{\mu}_k + (1 - \alpha_k)\mu_{L,k}.$$

Rearranging,

$$\hat{\mu}_k - \mu_k^{\mathrm{all}} = (1 - \alpha_k)(\hat{\mu}_k - \mu_{L,k}).$$

Taking norms and using the assumption gives

$$\|\hat{\mu}_k - \mu_k^{\mathrm{all}}\|_2 \le (1 - \alpha_k)\|\hat{\mu}_k - \mu_{L,k}\|_2 \le (1 - \alpha_k)\tau_k.$$

□

C.2 Lemma 3 and Proof

Lemma 3 (Sensitivity Budget Invariant). *Let PBMM (Algorithm 1) be run with per-institution budget $\varepsilon_k^{\max}$, failure probability δ, and noise scale σ_{DP}. We define the maximum admissible sensitivity*

$$S_k^{\max} := \frac{\varepsilon_k^{\max} \cdot \sigma_{\mathrm{DP}}}{\sqrt{2 \ln(1.25/\delta)}}. \tag{12}$$

Then PBMM guarantees, for every participating institution k:

*(i) **Budget Invariant.** At every step of the monotone radial expansion (Stage 3), the admitted set X_k^{sel} satisfies*

$$S_k := \left\|\mu(X_k^{\mathrm{sel}}) - \mu_k^{\mathrm{all}}\right\|_2 \le S_k^{\max} \tag{13}$$

*(ii) ϵ-δ-DP_{pool} **budget compliance for Pooled Mean Release.** The composed pipeline, i.e., PBMM selection followed by releasing $\tilde{\mu}(\mathcal{P}) = \mu(\mathcal{P}) + \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I)$, satisfies $(\varepsilon_k^{\max}, \delta)$-$DP_{pool}$ with respect to the in-domain neighbourhood of Definition 4 for each institution k.*

Proof. **(i)** *Base case.* After Stage 1, $X_k^{\mathrm{sel}} = \mathcal{M}_k$. By Lemma 1, $\|\hat{\mu}_k - \mu_k^{\mathrm{all}}\|_2 \le (1-\alpha_k)\tau_k$, and the hard-exclusion rule ($\mathcal{M}_k = \emptyset \Rightarrow$ institution excluded) ensures only institutions with a non-trivial matched core proceed. For any such institution, the initial sensitivity satisfies $S_k \le (1-\alpha_k)\tau_k$, which is within $S_k^{\max}$ whenever the matched core is non-negligible and the overlap radius τ_k is bounded (Lemma 1).

Inductive step. Suppose $S_k \le S_k^{\max}$ holds for the current admitted set X_k^{sel}. When Stage 3 considers admitting band $\mathcal{B}_b^{(k)}$, it computes the updated sensitivity $S_k' = \|\mu(X_k^{\mathrm{sel}} \cup \mathcal{B}_b^{(k)}) - \mu_k^{\mathrm{all}}\|_2$ and derives ε_k' via $\varepsilon = \frac{S\sqrt{2\ln(1.25/\delta)}}{\sigma_{\mathrm{DP}}}$. If $\varepsilon_k' > \varepsilon_k^{\max}$, the hard enforcement sets $\mathrm{score}(\mathcal{B}_b^{(k)}) = -\infty$, preventing admission. Therefore no band is ever admitted that would violate Eq. 13, and $S_k \le S_k^{\max}$ is maintained as a loop invariant throughout Stage 3.

(ii) By Part (i), the sensitivity of $\mu(X_k^{\mathrm{sel}})$ with respect to replacing X_k^{sel} with $\mathcal{D}_k$ (Definition 4) is bounded by $S_k^{\max}$. Plugging into the Gaussian mechanism ($\varepsilon = \frac{S\sqrt{2\ln(1.25/\delta)}}{\sigma_{\mathrm{DP}}}$): $\varepsilon_k = \frac{S_k^{\max}\sqrt{2\ln(1.25/\delta)}}{\sigma_{\mathrm{DP}}} = \varepsilon_k^{\max}$. Hence releasing $\tilde{\mu}(\mathcal{P}) = \mu(\mathcal{P}) + \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I)$ satisfies $(\varepsilon_k^{\max}, \delta)$-DP for institution k by the standard Gaussian mechanism theorem (Dwork & Roth, 2014). □

Remark. Lemma 3 establishes that PBMM provides a *formal* $(\varepsilon_k^{\max}, \delta)$-$DP_{pool}$ guarantee for the release of the pooled mean $\mu(\mathcal{P})$, consistent with the existing note (Appendix A.1) that no end-to-end DP of the final trained model is claimed. The anchor domain $\mathcal{D}_a$ is treated as a trusted coordinating institution (a standard assumption in federated protocols (Dwork & Roth, 2014)) and its privacy is outside the scope of the per-institution budget $\varepsilon_k^{\max}$.

C.3 Proposition 2 and Proof

Proposition 2 (**Pushout factorization of Stage 1 matching into feature space**). *Let $k \neq a$ be a non-anchor institution with matched subset $M_k \subseteq D_k$. Let $A := \{\hat{c}_c : c \in [C]\}$ be the anchor-centroid skeleton, and define $\gamma_k : M_k \to A$ by $\gamma_k(x) = \hat{c}_y$ for every $x \in M_k \cap D_k^{(y)}$. We form the pushout $Q_k := A \sqcup_{M_k} D_k$. Let $\Phi : D_k \to \mathbb{R}^{d_\Phi}$ be the encoder, and let there exist $g_k : A \to \mathbb{R}^{d_\Phi}$ defined by $g_k(\hat{c}_c) = \hat{c}_c^{\Phi}$.*

$$\begin{array}{ccc} M_k & \xrightarrow{\gamma_k} & A \\ {\scriptstyle i_k}\downarrow & & \downarrow{\scriptstyle j_k} \\ D_k & \xrightarrow[q_k]{} & Q_k := A \sqcup_{M_k} D_k \end{array}$$

Assume that Φ is consistent with the anchor skeleton on M_k in the sense that $\Phi(x) = \hat{c}_y^{\Phi}$ for every $x \in M_k \cap D_k^{(y)}$, or more generally that $\Phi|_{M_k} = g_k \circ \gamma_k$. Then there exists a unique map $\bar{\Phi}_k : Q_k \to \mathbb{R}^{d_\Phi}$ such that $\bar{\Phi}_k \circ q_k = \Phi, \quad \bar{\Phi}_k \circ j_k = g_k$.

$$\begin{array}{ccc} M_k & \xrightarrow{\gamma_k} & A \\ {\scriptstyle i_k}\downarrow & & \downarrow{\scriptstyle g_k} \\ D_k & \xrightarrow[\Phi]{} & \mathbb{R}^{d_\Phi} \end{array} \Longrightarrow \textit{there exists a unique map } \bar{\Phi}_k : Q_k \to \mathbb{R}^{d_\Phi}$$

Consequently, every matched sample and its assigned anchor class-side have the same image under the induced map $\bar{\Phi}_k$, i.e. $\bar{\Phi}_k(q_k(x)) = \bar{\Phi}_k(j_k(\hat{c}_y))$ for every $x \in M_k \cap D_k^{(y)}$. In particular, matching relation is preserved after passing from data space to feature space.

Proof. Pushouts exist in Set Mac Lane & MacLane (1971), so $\mathcal{Q}_k$ is well-defined by the universal property alone.

Remark. The pushout $\mathcal{Q}_k = A \sqcup_{\mathcal{M}_k} \mathcal{D}_k$ can be constructed explicitly as $(\mathcal{D}_k \sqcup A)/\!\sim$, where $\sim$ is the equivalence relation generated by $x \sim \gamma_k(x)$ for each $x \in \mathcal{M}_k$, and then taking the transitive closure.

The transitive closure is necessary here because γ_k is *not* injective: all matched samples from the same class y are identified with the single centroid $\hat{c}_y$, so distinct samples x, x' with $\gamma_k(x) = \gamma_k(x') = \hat{c}_y$ must also be identified with each other.

Since $\Phi|_{M_k} = g_k \circ \gamma_k$, the maps $\Phi : D_k \to \mathbb{R}^{d_\Phi}$ and $g_k : A \to \mathbb{R}^{d_\Phi}$ agree on the gluing data M_k. Therefore, by the universal property of the pushout, there exists a unique map $\bar{\Phi}_k : Q_k \to \mathbb{R}^{d_\Phi}$ such that $\bar{\Phi}_k \circ q_k = \Phi, \quad \bar{\Phi}_k \circ j_k = g_k$ [5].

Now let $x \in M_k \cap D_k^{(y)}$. Since $q_k(x) = j_k(\hat{c}_y)$ in the pushout, applying $\bar{\Phi}_k$ gives $\bar{\Phi}_k(q_k(x)) = \bar{\Phi}_k(j_k(\hat{c}_y))$. Hence the matched sample and its assigned anchor class-side have the same image under the induced feature-space map. Therefore the Stage 1 matching relation is preserved after transport to feature space. □

C.4 Assumption 1, Lemma 4 and Proof

Assumption 1 (Radial concentration and practical monotonicity). *For each institution k, the feature distribution is approximately radially concentrated around its institution mean μ_k^{all}. In particular, for the radial direction u_y used in $\mathcal{C}_\dagger$, samples farther along the ray $x(r) = \mu_k^{\text{all}} + ru_y, \quad r \geq 0$, are farther from the typical in-domain region of institution k. Equivalently, for $0 \leq r_1 < r_2 < r_3$, $\|x(r_1) - \mu_k^{\text{all}}\|_2 < \|x(r_2) - \mu_k^{\text{all}}\|_2 < \|x(r_3) - \mu_k^{\text{all}}\|_2$. This assumption is practical for Gaussian, sub-Gaussian, and radially log-concave feature distributions, where empirical radial shells concentrate around their population radii as n_k grows. Thus the inner/mid/outer radial bands used by PBMM inherit a stable privacy-cost ordering in finite samples, rather than depending on pathological individual points.*

Lemma 4 (Functoriality of utility and privacy summaries). *Let $\mathcal{C}_y$ be the forward-translation category whose objects are points $x(r) = \mu_k^{\text{all}} + ru_y, \quad r \geq 0$, and whose morphisms are translations $f_t : x(r) \to x(r+t), \quad t \geq 0$. Composition is defined by $f_s \circ f_t = f_{s+t}$, and the identity morphism on $x(r)$ is $f_0 : x(r) \to x(r)$.*

Let's define $F_{\text{util}}(x) = \|x - \hat{c}_{1-y}\|_2^2 - \|x - \hat{c}_y\|_2^2$, and $F_{\text{priv}}(x) = -\|x - \mu_k^{\text{all}}\|_2^2$. For every morphism $f_t : x(r) \to x(r+t)$, we define $F_{\text{util}}(f_t) : F_{\text{util}}(x(r)) \to F_{\text{util}}(x(r+t))$ to be the unique order morphism in $(\mathbb{R}, \leq)$ whenever $F_{\text{util}}(x(r)) \leq F_{\text{util}}(x(r+t))$, and we also define $F_{\text{priv}}(f_t) : F_{\text{priv}}(x(r)) \to F_{\text{priv}}(x(r+t))$ to be the unique order morphism in $(\mathbb{R}, \geq)$ whenever $F_{\text{priv}}(x(r)) \geq F_{\text{priv}}(x(r+t))$. Then $F_{\text{util}} : \mathcal{C}_y \to (\mathbb{R}, \leq)$ and $F_{\text{priv}} : \mathcal{C}_y \to (\mathbb{R}, \geq)$ are functors on the subcategory of admissible translations satisfying the above monotonicity conditions.

Proof. A functor must specify an object map, a morphism map, preserve identity morphisms, and preserve composition.

First, the object maps are well-defined because both $F_{\text{util}}(x) = \|x - \hat{c}_{1-y}\|_2^2 - \|x - \hat{c}_y\|_2^2$ and $F_{\text{priv}}(x) = -\|x - \mu_k^{\text{all}}\|_2^2$ assign a real number to every object $x \in \mathcal{C}_y$.

Second, consider a morphism $f_t : x(r) \to x(r+t)$. The target categories $(\mathbb{R}, \leq)$ and $(\mathbb{R}, \geq)$ are poset categories: between two objects there exists a unique morphism exactly when the corresponding order relation holds. Therefore, if $F_{\text{util}}(x(r)) \leq F_{\text{util}}(x(r+t))$, then there is a unique morphism $F_{\text{util}}(x(r)) \to F_{\text{util}}(x(r+t))$ in $(\mathbb{R}, \leq)$. Similarly, if $F_{\text{priv}}(x(r)) \geq F_{\text{priv}}(x(r+t))$, then there is a unique morphism $F_{\text{priv}}(x(r)) \to F_{\text{priv}}(x(r+t))$ in $(\mathbb{R}, \geq)$. Hence the morphism maps are well-defined on admissible translations.

[5] *The non-triviality lies in what happens to the unmatched samples $\mathcal{D}_k \setminus M_k$: the assumption only specifies Φ on M_k, yet the universal property forces a unique, globally coherent extension $\bar{\Phi}_k$ that covers the entire dataset including unmatched samples, with no ambiguity and no additional choices. This matters for Stages 2 and 3: the radial bands, privacy-tempered centroids, and monotone expansion are all built around $\hat{\mu}_k = \mu(M_k)$ as the certified core geometry. Proposition 2 guarantees that this core geometry, and the anchor-consistent ordering it induces, is the same object in feature space as in data space. So the utility-privacy tension of Theorem 1 and the DP-cost ordering of Proposition 1 carry over to Algorithm 1 without modification, making the heterogeneous feature-space instantiation theoretically grounded rather than a heuristic transposition of the true unconfounded causal relation in data-space analysis.*

Third, identities are preserved. The identity morphism in $\mathcal{C}_y$ is $f_0 : x(r) \to x(r)$. Since $F_{\text{util}}(x(r)) = F_{\text{util}}(x(r))$ and $F_{\text{priv}}(x(r)) = F_{\text{priv}}(x(r))$, the induced morphisms are exactly the identity morphisms on $F_{\text{util}}(x(r))$ and $F_{\text{priv}}(x(r))$, respectively.

Fourth, composition is preserved. Let $x(r) \xrightarrow{f_t} x(r+t) \xrightarrow{f_s} x(r+t+s)$ be two composable admissible morphisms. In $\mathcal{C}_y$, $f_s \circ f_t = f_{s+t}$. Because the target categories are poset categories, the morphism from $F(x(r))$ to $F(x(r+t+s))$, when it exists, is unique. Therefore $F(f_s \circ f_t) = F(f_{s+t}) = F(f_s) \circ F(f_t)$, for $F \in \{F_{\text{util}}, F_{\text{priv}}\}$. Thus both maps preserve composition.

It remains to justify the privacy monotonicity used for F_{priv}. By construction, $x(r) = \mu_k^{\text{all}} + ru_y$, with $\|u_y\|_2 = 1$. Hence $\|x(r) - \mu_k^{\text{all}}\|_2^2 = \|ru_y\|_2^2 = r^2$. Therefore $F_{\text{priv}}(x(r)) = -r^2$. If $0 \le r \le r+t \le r+t+s$, then $-r^2 \ge -(r+t)^2 \ge -(r+t+s)^2$, so F_{priv} is order-preserving as a map into $(\mathbb{R}, \ge)$. Under the radial concentration assumption, this radial ordering matches the empirical inner-to-outer privacy ordering used by PBMM. Therefore F_{util} and F_{priv} are valid functors on the admissible forward-translation category. □

C.5 Proof of Theorem 1

Theorem 1 (**Functorial incompatibility under shift**). *Let k be an unharmonized institution, and fix a class y such that $v_y := \hat{c}_y - \hat{c}_{1-y} \neq 0$. Then there exist composable morphisms $x \xrightarrow{f} x' \xrightarrow{g} x''$ in $\mathcal{C}_y$ such that F_{util} maps them to valid order-preserving morphisms f', g' in $(\mathbb{R}, \le)$ satisfying $U(x) < U(x') < U(x'')$, whereas F_{priv} reverses the order, yielding $V(x) > V(x') > V(x'')$.*

Proof. Under Assumption 1, Lemma 4 establishes that F_{util} and F_{priv} define valid functors over admissible translations and induce a consistent monotone ordering along radial directions.

By construction, the morphisms of $\mathcal{C}_y$ are translations. Fix a class y with $v_y = \hat{c}_y - \hat{c}_{1-y} \neq 0$, and define the unit direction $u_y := v_y / \|v_y\|_2$. Consider the objects $x = \mu_k^{\text{all}} + ru_y$, $x' = \mu_k^{\text{all}} + r'u_y$, and $x'' = \mu_k^{\text{all}} + r''u_y$ for $0 \le r < r' < r''$. Let $f : x \to x'$ and $g : x' \to x''$ be the corresponding translation morphisms. Since translation addition is transitive, $g \circ f : x \to x''$ is a valid composable morphism. For arbitrary $r \ge 0$, let $z := x(r) = \mu_k^{\text{all}} + ru_y$.

First, evaluating the utility object map $U(z) = \|z - \hat{c}_{1-y}\|_2^2 - \|z - \hat{c}_y\|_2^2$ along this path yields:

$$U(x(r)) = 2\langle \mu_k^{\text{all}}, v_y \rangle + 2r\|v_y\|_2 + C_y \tag{14}$$

where $C_y = \|\hat{c}_{1-y}\|_2^2 - \|\hat{c}_y\|_2^2$. Since $\|v_y\|_2 > 0$, this is strictly increasing, guaranteeing $U(x) < U(x') < U(x'')$. Consequently, unique morphisms $f' : U(x) \to U(x')$ and $g' : U(x') \to U(x'')$ exist in the poset category $(\mathbb{R}, \le)$. The functor F_{util} maps $f \mapsto f'$ and $g \mapsto g'$. Because there is at most one morphism between any two objects in a poset, the composition $g' \circ f'$ is the unique morphism $U(x) \to U(x'')$, which identically matches $F_{\text{util}}(g \circ f)$. Identity is preserved as $U(x) \le U(x)$ yields $\text{id}_{U(x)}$. Thus, the functorial axioms of identity and composition hold for utility.

Second, evaluating the privacy object map $V(z) = -\|z - \mu_k^{\text{all}}\|_2^2$ gives $V(x(r)) = -r^2$. For $r < r' < r''$, we have:

$$V(x) > V(x') > V(x'') \tag{15}$$

Therefore, the mapped morphisms f', g' do not exist in $(\mathbb{R}, \le)$, but rather strictly in the opposite category $(\mathbb{R}, \ge)$. Thus, any sequence of morphisms f, g that improves matching utility structurally reverses the relation required for in-domain privacy. □

D Algorithmic explanation and Ablation studies

D.1 Algorithmic explanation of PBMM for two distinct setups

An algorithmic summarization of the proposed PBMM is shown in Algorithm 1.

D.2 Ablation Studies

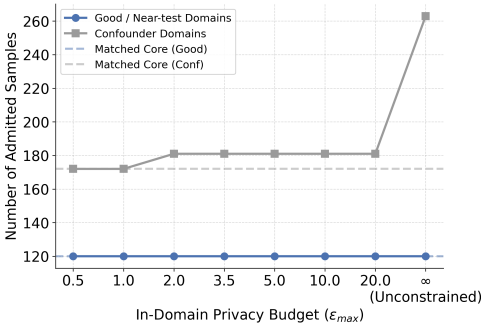


Figure 16: **Effect of $\varepsilon_k^{\max}$ on the number of samples selected.** Increasing the privacy budget ($\varepsilon_k^{\max}$) leads to increase in the number of selected samples.

Effect of Privacy budget. We use the synthetic setup (Sec. B.1) to study the effects of varying the privacy budget $\varepsilon_k^{\max}$. By default, $\varepsilon_k^{\max}$ was set to ∞. Restricting it to smaller values leads to lesser selection of samples. For the domains that are unconfounded, i.e., the ones that are closer to the test domain, Matching ($\varepsilon_k^{\max} = \infty$) and PBMM mimic each other. For confounded domains, increasing the privacy leads to increase in the inclusion of number of samples for PBMM compared to matching. This is a direct indication of the contrasting objectives of utility and privacy as seen in Theorem 1.

Ablation on privacy-tempering kernels. In addition to the default Gaussian privacy-tempering rule, we study several alternative radial kernels for the online PBMM-style admission step. Let $r(x) = \|x - \hat{\mu}_k\|$ denote the distance of a candidate sample from the running in-domain core mean, and let $\tau_k > 0$ be the corresponding scale parameter. We instantiate the privacy weight $w_k(x)$ using five choices: *Gaussian* (used as the default in all experimental settings) $w_k^{\text{Gauss}}(x) = \exp\left(-\frac{r(x)^2}{2\tau_k^2}\right)$, *Laplacian* $w_k^{\text{Lap}}(x) = \exp\left(-\frac{r(x)}{\tau_k}\right)$, *generalized Gaussian* $w_k^{\text{GenGauss}}(x) = \exp\left(-\left(\frac{r(x)}{\tau_k}\right)^p\right), \quad p > 0$ which recovers the Laplacian case at $p = 1$ and the Gaussian case at $p = 2$, *Staircase* (Geng et al., 2015) $w_k^{\text{Stair}}(x) = \alpha_j \quad \text{if } r(x) \in [jh, (j+1)h), \quad \alpha_{j+1} = \rho\,\alpha_j,\ 0 < \rho < 1$ which yields a shell-wise piecewise-constant radial weighting inspired by staircase mechanisms Geng et al. (2015), and *Compact* $w_k^{\text{Comp}}(x) = \max\left(0,\ 1 - \left(\frac{r(x)}{h_k}\right)^2\right)$ which imposes a hard cutoff beyond radius h_k. In all cases, the weighted centroid is formed as $\tilde{c}_y^{(k)} = \frac{\sum_{x \in D_k^{(y)}} w_k(x)\, x}{\sum_{x \in D_k^{(y)}} w_k(x)}$ and is then used in the same privacy-tempered margin as in the main method. This ablation isolates the effect of the radial decay profile itself: Gaussian weights enforce the strongest concentration around the in-domain core, Laplacian and generalized-Gaussian kernels interpolate between sharper central emphasis and heavier-tail retention, Staircase yields shell-wise discrete attenuation, and Compact strictly truncates far-tail contributions. As shown in Table 2, the Gaussian and Laplacian variants achieve the strongest downstream utility, whereas generalized Gaussian and Compact are substantially more conservative and lead to lower AUC. In contrast, the Staircase kernel incurs the largest middle- and outer-band privacy costs, suggesting that shell-wise decay is less effective than smooth radial decay for suppressing tail admissions in this setting.

Table 2: **Ablation of privacy-tempering kernels.** For each radial band, we report in-domain sensitivity and privacy cost as S/ϵ. Lower values indicate lower privacy exposure.

Kernel	**Inner** (S/ϵ)	**Mid** (S/ϵ)	**Outer** (S/ϵ)	**AUC**
Gaussian	0.010 / 0.203	0.052 / 0.110	0.147 / 3.118	83.54
generalized Gaussian	0.007 / 0.152	0.005 / 0.103	0.027 / 0.562	68.98
Laplacian	0.006 / 0.134	0.072 / 1.518	0.132 / 2.805	83.54
Staircase	0.055 / 1.163	0.188 / 3.980	0.292 / 6.180	78.54
Compact	0.009 / 0.198	0.071 / 1.500	0.147 / 3.110	66.60

Multi-class Scalability and Privacy-Utility Trade-offs. We extend our evaluation to multi-class classification to investigate the scalability and robustness of PBMM. We vary the number of classes $C \in \{2, 3, 4, 8, 10\}$ and isolate both the in-domain privacy cost ε and downstream test accuracy across three distinct radial bands (I_{inner}, I_{mid}, and I_{outer}). To accommodate label spaces where $C > 2$, the binary margin is naturally generalized: a candidate sample's inclusion is governed by the relative distance between its own-class centroid and its nearest out-of-class centroid, effectively enforcing a strict multi-class confidence boundary. As illustrated in Figure 17, baseline methods struggle to balance the natural tension between harmonization and privacy as the label space grows. Naive Pooling and standard Matching incur severe, scaling privacy costs as C increases (particularly in the Mid and Outer bands, where domain-shifted tail samples are most prevalent). Notably, all methods exhibit a downward trend in absolute test accuracy as C increases; this decay is expected as the increased complexity of the decision boundaries requires higher sample density per class to maintain performance. Subsampling successfully restricts privacy leakage but suffers catastrophic utility degradation in

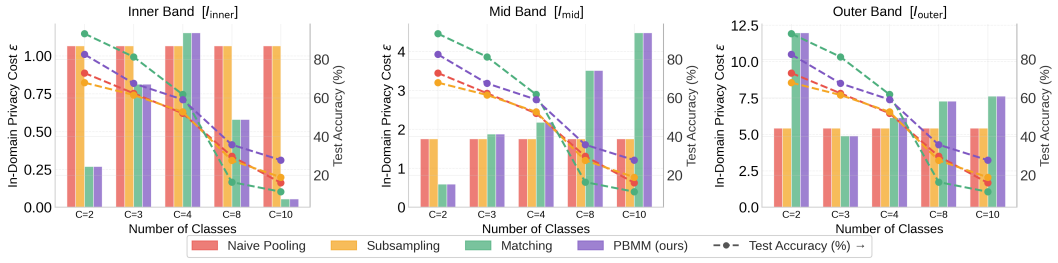


Figure 17: **Multi-class ablation.** Evaluating in-domain privacy cost ε (bars) and downstream test accuracy (dashed lines) across three radial bands (I_{inner}, I_{mid}, I_{outer}) as the number of classes C varies from 2 to 10. PBMM successfully bounds privacy leakage in the tail bands while maintaining competitive utility compared to unconstrained baselines.

these high-C regimes. Conversely, PBMM strictly controls the privacy cost ε across all bands while maintaining a competitive test accuracy that tightly tracks the upper-bound baselines. This ablation confirms that PBMM's radial decay profile effectively suppresses risky tail admissions, ensuring that privacy exposure remains bounded even as the task complexity scales.

Ablation on Band Granularity (Eq. 7). We evaluate the sensitivity of PBMM to the number of radial bands n, which governs the granularity of the privacy-tempered margin. To assess this, we measure *Monotonicity* (the preservation of the theoretical cost ordering $\varepsilon_b \leq \varepsilon_{b+1}$) and *Discriminability* (the standard deviation of privacy costs across bands, capturing the resolution of the privacy signal). As detailed in Table 3, using a single band fails to capture any distributional variance. While increasing the bands to $n = 5$ yields a sharp increase in discriminability (4.6351), it comes at the direct cost of monotonicity, which decays steadily ($0.8750 \rightarrow 0.7778$) as finer partitions introduce sample sparsity and violate the cost ordering. For PBMM, $n = 3$ emerges as the critical stabilization point: it restores monotonicity to a robust peak of 0.9167 while providing a meaningful discriminative signal (1.7604), cementing three bands as the optimal resolution for balancing privacy stability and utility (as highlighted in Theorem 1).

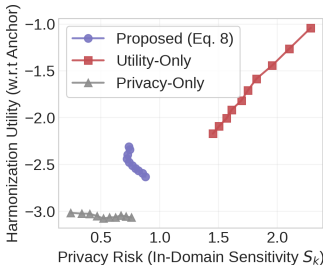


Figure 18: **Ablation of the scoring function components.** The proposed joint score (Avg. Utility: -1.98, Avg. Sensitivity: 0.94) achieves a superior trade-off compared to Utility-Only (Avg. Utility: -0.56, Avg. Sensitivity: 3.24) and Privacy-Only (Avg. Utility: -3.43, Avg. Sensitivity: 0.69).

Ablation on score (Eq. 4. To assess the role of privacy-tempered radial decay in our admission policy, we compare against two baselines: *Utility-Only* (standard matching without radial bias) and *Privacy-Only* (selection based solely on in-domain centrality). As shown in Fig. 18, Utility-Only achieves strong distributional alignment but incurs prohibitive sensitivity ($S_k \approx 3.24$), quickly exhausting the ϵ-budget on a few tail samples. In contrast, Privacy-Only minimizes risk ($S_k \approx 0.69$) but yields poor harmonization (Avg. Utility ≈ -3.43). Our joint scoring rule (Eq. 8) recovers a Pareto-optimal trade-off, maintaining competitive alignment (Avg. Utility ≈ -1.98) while reducing sensitivity by over 70% relative to greedy matching. This demonstrates the importance of exponential radial decay. This demonstrates that exponential radial decay is essential for prioritizing privacy-safe core samples without sacrificing alignment to the anchor domain.

Table 3: **Band Granularity Ablation.** $n = 3$ (our choice) provides the optimal balance, maximizing the stability of the privacy ordering (Monotonicity) before finer partitions cause it to decay.

Method	Bands	Monot. ↑	Discrim. ↑	Marg. Gain
Matching	1	1.0000	0.0000	0.0000
	2	1.0000	5.0629	5.0629
	3	**0.9167**	**5.4937**	**0.4308**
	5	0.8750	5.2824	-0.2113
	7	0.8611	6.1049	0.8225
	10	0.8333	5.5142	-0.5907
PBMM (Ours)	1	1.0000	0.0000	0.0000
	2	0.8333	1.1648	1.1648
	3	**0.9167**	**1.7604**	**0.5955**
	5	0.8750	4.6351	2.8747
	7	0.8333	4.7021	0.0670
	10	0.7778	5.7149	1.0127

Empirical adversaries. We instantiate two adversaries that operationalize the threat framing of Sec. 3 in the trained encoder's feature space, comparing Matching against PBMM on the homogeneous ABIDE curriculum (Sec. B.2). *(a)* targets distributional identity leakage: for each non-anchor site k, the adversary observes $\mu(X_k^{\text{sel}})$, and by Lem. 2 the sufficient statistic controlling its optimal Neyman-Pearson advantage is $S_k = \|\mu(X_k^{\text{sel}}) - \mu_k^{\text{all}}\|_2$. *(b)* targets per-record tail-membership inference: knowing the anchor's per-class centroids and each

candidate's embedding, the adversary scores every train-pool record with the matching margin (Eq. 1) and predicts which records were admitted. *(b)* is restricted to the aggregator's candidate set (the train pool) so the AUC isolates the selection-level signal.

Fig. 19 decomposes the exposure exactly along the paper's framing (Sec. 1): *what leaks under matching-style pooling is not a patient record but the shape of a hospital's population.* On (a), Matching's median S_k ($\approx$1.9) is roughly twice PBMM's ($\approx$ 1.0), with a heavy upper tail to $\sim$ 3.9; PBMM's radial admission compresses this distribution and pulls mass toward the LLN-consistent low-S_k region. On (b), both methods sit within a narrow band just above chance (median AUC ≈ 0.58, upper whisker ~ 0.74). Under a realistic adversary whose knowledge is the released encoder and anchor centroids (rather than the exact online-EMA state at selection time) per-record tail-membership inference is not the primary exposure surface of matching-style pooling; distributional identity leakage is. (a) is the exposure PBMM was designed to control, and (b) confirms that PBMM does not open a compensating per-record channel: Matching's structurally distributional failure is corrected without inducing a record-level one.

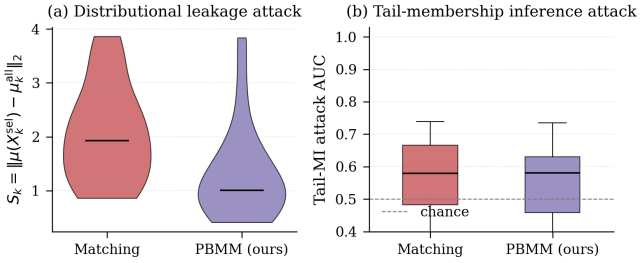


Figure 19: **Empirical privacy attacks on homogeneous ABIDE, in ResNet-18 feature space, Matching vs. PBMM. (a) Distributional leakage attack:** per-site in-domain sensitivity $S_k = \|\mu(X_k^{\mathrm{sel}}) - \mu_k^{\mathrm{all}}\|_2$, the sufficient statistic of the optimal Neyman–Pearson adversary against a released contribution (Lem. 2). **(b) Tail-membership inference attack:** per-site AUC of the anchor-based matching margin (Eq. 1) against the true selection mask, on the aggregator's train pool. Under a realistic adversary with access only to the released encoder and anchor centroids.

E LIMITATIONS AND FUTURE SCOPE.

By preferring short displacements, PBMM admits more from an institution's typical region and fewer of its rare presentations, which sit in the outer shells — a cost that is not incidental but the content of Thm. 1. Stage 1 is matching unchanged, so anchor-consistent rare samples still enter through the matched core which accounts for it but filtering does indeed leave out some rare presentations (outliers) in clinical sense.

Algorithm 1 PBMM

Require: Domains $\{\mathcal{D}_k\}_{k=1}^{K}$, encoder Φ, anchor a, per-institution budgets $\{\varepsilon_k^{\max}\}$, failure probability δ, noise scale σ_{DP}
Ensure: Privacy-aware pooled dataset $\mathcal{P}$

// Stage 1: Matched core in feature space
$\hat{c}_c \leftarrow \mathrm{mean}\Big(\Phi\Big(\mathcal{D}_a^{(c)}\Big)\Big)$ for each $c \in [C]$ ▷ Initialise centroids from anchor embeddings
$\mathcal{X}_{\mathrm{core}} \leftarrow \mathcal{D}_a$
for $k = 1, \ldots, K,\ k \neq a$ **do**
  $\mathbf{z}_i \leftarrow \Phi(x_i)$ for each $x_i \in \mathcal{D}_k$ ▷ Encode domain k with shared Φ
  $\mathcal{M}_k \leftarrow \big\{x_i \in \mathcal{D}_k : \|\mathbf{z}_i - \hat{c}_{y_i}\| < \|\mathbf{z}_i - \hat{c}_j\|\ \forall\, j \neq y_i\big\}$ ▷ Eq. 1 in $\mathbb{R}^{d_\Phi}$
  $\mathcal{X}_{\mathrm{core}} \leftarrow \mathcal{X}_{\mathrm{core}} \cup \mathcal{M}_k$
  $\hat{c}_c \leftarrow \frac{1}{2}\Big(\hat{c}_c + \bar{\mathbf{z}}_c^{(k)}\Big)$ for each c, where $\bar{\mathbf{z}}_c^{(k)}$ is the mean embedding of $\{x_i \in \mathcal{M}_k : y_i = c\}$ ▷ Online EMA; hard-exclude k if $\mathcal{M}_k = \emptyset$
end for

// Stage 2: Privacy-tempered expansion in feature space
for each k with $|\mathcal{M}_k| > 0,\ k \neq a$ **do**
  $\hat{\boldsymbol{\mu}}_k \leftarrow |\mathcal{M}_k|^{-1} \sum_{x_i \in \mathcal{M}_k} \Phi(x_i), \quad \hat{\sigma}_k^2 \leftarrow |\mathcal{M}_k|^{-1} \sum_{x_i \in \mathcal{M}_k} \|\Phi(x_i) - \hat{\boldsymbol{\mu}}_k\|^2$ ▷ Core statistics in $\mathbb{R}^{d_\Phi}$; Lemma 1
  $R_k \leftarrow \max_{x_i \in \mathcal{D}_k \setminus \mathcal{M}_k} \|\Phi(x_i) - \hat{\boldsymbol{\mu}}_k\|$ ▷ Normalisation radius in $\mathbb{R}^{d_\Phi}$
  $X_k^{\mathrm{sel}} \leftarrow \mathcal{M}_k$ ▷ Initialise admitted set for cumulative budget tracking
  for each class $y \in [C]$ **do**
    $\tilde{c}_y^{(k)} \leftarrow \dfrac{\sum_{x_i \in \mathcal{D}_k^{(y)}} w_k(x_i)\, \Phi(x_i)}{\sum_{x_i \in \mathcal{D}_k^{(y)}} w_k(x_i)}, \quad w_k(x_i) = \exp\Big(-\frac{\|\Phi(x_i) - \hat{\boldsymbol{\mu}}_k\|^2}{2\hat{\sigma}_k^2}\Big)$ ▷ Class-specific centroid in $\mathbb{R}^{d_\Phi}$; Eq. 5
  end for
  for $x_i \in \mathcal{D}_k \setminus \mathcal{M}_k$ **do**
    $m_{\mathrm{priv}}(x_i) \leftarrow R_k^{-2}\Big[\big\|\Phi(x_i) - \hat{c}_{\arg\min_{j \neq y_i} \|\Phi(x_i) - \hat{c}_j\|}\big\|^2 - \|\Phi(x_i) - \tilde{c}_{y_i}^{(k)}\|^2\Big]$ ▷ Eq. 6 in $\mathbb{R}^{d_\Phi}$
  end for
  Partition $\{x_i : m_{\mathrm{priv}}(x_i) > 0\}$ into bands $\mathcal{B}_{\mathrm{in}}^{(k)}$, $\mathcal{B}_{\mathrm{mid}}^{(k)}$, $\mathcal{B}_{\mathrm{out}}^{(k)}$ by $\|\Phi(x_i) - \hat{\boldsymbol{\mu}}_k\|/R_k \in I_b$ ▷ Eq. 7; radial order preserved by Prop. 1
  for $b \in \{\mathrm{in}, \mathrm{mid}, \mathrm{out}\}$ **do**
    if $|\mathcal{B}_b^{(k)}| = 0$ **then**
      $\mathrm{score}(\mathcal{B}_b^{(k)}) \leftarrow 0$ ▷ Empty band auto-exhausted; does not block expansion
    else
      $\mathrm{score}(\mathcal{B}_b^{(k)}) \leftarrow \log\big(|\mathcal{B}_b^{(k)}| + 1\big) + |\mathcal{B}_b^{(k)}|^{-1} \sum_{x_i \in \mathcal{B}_b^{(k)}} m_{\mathrm{priv}}(x_i)$ ▷ Eq. 8
    end if
    // Sensitivity computed locally by institution k; only scalar ε_k is communicated centrally
    $S_k \leftarrow \Big\|\mu\Big(X_k^{\mathrm{sel}} \cup \mathcal{B}_b^{(k)}\Big) - \mu_k^{\mathrm{all}}\Big\|_2$ ▷ Cumulative sensitivity; μ_k^{all} known only to institution k
    Compute ε_k via $\varepsilon = \frac{S\sqrt{2\ln(1.25/\delta)}}{\sigma_{\mathrm{DP}}}$ using S_k
    if $\varepsilon_k > \varepsilon_k^{\max}$ **then**
      $\mathrm{score}(\mathcal{B}_b^{(k)}) \leftarrow -\infty$ ▷ Hard budget enforcement
    end if
  end for
end for

// Stage 3: Monotone radial expansion
$\mathcal{P} \leftarrow \mathcal{X}_{\mathrm{core}}$
while $\exists\ \mathcal{B}_b^{(k)}$ with $\mathrm{score}(\mathcal{B}_b^{(k)}) > 0$ **do**
  $\mathcal{B}^\star \leftarrow \arg\max_{k,b}\ \mathrm{score}(\mathcal{B}_b^{(k)})$
    subject to: $b = \min\big\{b' : \mathcal{B}_{b'}^{(k)}$ not yet admitted and $|\mathcal{B}_{b'}^{(k)}| > 0\big\}$ ▷ Innermost non-empty available band; empty bands skipped, preserving radial order (Prop. 1)
  $X_k^{\mathrm{sel}} \leftarrow X_k^{\mathrm{sel}} \cup \mathcal{B}^\star$ ▷ Update admitted set for cumulative budget tracking
  $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{B}^\star$
end while
return $\mathcal{P}$