PROTECTING THE GROUP, NOT JUST ITS MEMBERS: FEASIBLE DIFFERENTIAL PRIVACY FOR GROUPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large amounts of data are required to infer robust statistics or train large machine-learned models. Not only a public release of such data can reveal sensitive information of the individuals, but such information can also be inferred from a statistical estimator or learned predictor. Differential privacy (DP) provides strong guarantees for protecting individual records; however, these guarantees do not prevent sensitive properties of groups from being inferred. In many applications including healthcare, protecting the characteristics of a subpopulation within an institution can be as important as protecting individual records, for example disease prevalence within a demographic group at a hospital. We introduce *Differentially Private Attributes of Groups (DPAG)*, a framework that treats such groups as the atomic units of privacy. Rather than obtaining group protection by scaling an individual-level privacy guarantee with group size, DPAG directly bounds and perturbs each protected group's contribution. We develop mechanisms for centralized learning and extend the framework to cross-silo federated learning that is regularly used in healthcare scenarios to transfer institutional knowledge for machine learning without sharing the underlying data. We evaluate DPAG on centralized and federated clinical data, comparing direct group-level protection with standard DP mechanisms and conventional group-privacy approaches. Our results show that **DPAG can make group-level protection feasible without a privacy cost that grows with the number of individuals in the group, while retaining useful model utility**.

1 INTRODUCTION

Modern machine learning increasingly relies on data collected across institutions, populations, and demographic groups. In domains such as healthcare or finance, such collections of data can support model training that would be difficult within any single organization. Collaboration, however, also creates privacy risks that go beyond the disclosure of individual records. For example, a hospital may want to protect not only whether a particular patient participated in training, but also sensitive characteristics of *groups of people*, such as disease prevalence within a demographic group or differences in clinical outcomes across subpopulations.

Differential privacy (DP) provides a rigorous and widely used framework for limiting what can be learned about individuals from released statistics or trained models (Dwork, 2006; Dwork et al., 2006; Dwork & Roth, 2014). In machine learning, methods such as DP-SGD bound the influence of individual training examples and add calibrated noise during optimization (Abadi et al., 2016). Differential privacy has also become an important tool in federated learning, where multiple clients collaboratively train a model without directly sharing their raw data (McMahan et al., 2017; 2018).

Yet, protecting individuals does not necessarily protect sensitive properties of their associated groups. A population-level characteristic may be strongly supported by many records and therefore remain learnable even when the influence of each individual is limited. Standard DP does provide a group-privacy guarantee for changes affecting several individuals, but this guarantee degrades with the number of individuals in the group, often making it prohibitively costly and impossible to apply in larger practical settings. This work considers a different privacy objective, which is *directly* protecting predefined groups as the units of privacy. We introduce *Differentially Private Attributes of Groups (DPAG)*, where a protected unit is a combination of group and institution – a cell – such

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.

as demographic group within a particular hospital. Rather than asking whether the contribution of one person can be detected, we ask whether the output of the learning procedure reveals sensitive characteristics of an *entire protected cell*. This enables our definition of privacy to match applications in which population characteristics themselves are confidential.

Here, we suggest that the contribution of each protected cell must be explicitly bounded before information from different cells is combined, which is achieved through calibrated noise for each protected cell. The resulting framework retains the familiar semantics and composition properties of differential privacy while changing the privacy unit to reflect the protection of group-level information. This perspective is particularly natural for cross-silo federated learning. Considering several hospitals collaboratively training a model, client-level privacy may protect the participation of an entire hospital, while individual-level privacy protects single patients. Neither guarantee directly captures an intermediate privacy objective such as protecting the characteristics of a particular ethnicity, sex, or age group within one hospital. DPAG is designed for precisely this intermediate level of granularity.

In experiments, we study DPAG in centralized and federated learning settings, including intersectional demographic groups in tabular clinical data and federated learning on medical images. Across these experiments, we compare direct cell-level protection with standard differentially private baselines and with mechanisms that attempt to obtain equivalent protection through conventional group-privacy scaling. The results illustrate that DPAG enables group-level protection for large data, avoiding the privacy cost of individual-level DP growing linearly with number of individuals in a group.

In summary, our main contributions are:

- We formulate DPAG, a differential-privacy framework in which predefined groups or institution–group cells are treated as the atomic privacy units.
- We develop centralized and federated learning mechanisms that bound contributions directly at the cell level, avoiding privacy cost that grows with the number of individuals in the cell.
- We evaluate the approach on tabular and medical-imaging tasks, illustrating the privacy–utility trade-offs of direct group protection in comparison to conventional DP baselines.

2 Related Work

Differential privacy. Differential privacy (DP) provides a worst-case guarantee that the output of a randomized mechanism changes only slightly when one privacy unit is modified (Dwork, 2006; Dwork et al., 2006; Dwork & Roth, 2014). In its standard form, that unit is an individual record. DP mechanisms typically bound the sensitivity of a statistic and add calibrated Laplace or Gaussian noise Balle & Wang (2018). In machine learning, DP-SGD applies this principle by clipping per-example gradients and perturbing their aggregate (Abadi et al., 2016). Repeated releases can be analyzed using tighter composition frameworks such as Rényi differential privacy (Mironov, 2017). Agarwal et al. (2025) considered differential privacy when each individual contributes multiple samples to the dataset. Other lines of works have studied alternative targets as privacy, for example only sample labels being private (Busa-Fekete et al., 2023; Ghazi et al., 2024; Jiang et al., 2024b),

A standard consequence of record-level DP is *group privacy*: if a mechanism protects one individual, it also provides a weaker guarantee when multiple records change simultaneously (Dwork & Roth, 2014). However, the privacy loss grows with the number of changed individuals. This makes classical group-privacy conversion unattractive when the desired protected object is a large demographic or institutional subgroup. DPAG instead follows the same principle of user-level DP by defining the desired object directly as the privacy unit and bounding its contribution at that level (McMahan et al., 2018; Kato et al., 2024; Levy et al., 2021; Asi et al., 2022). In our setting, this unit is an intersectional cell such as a hospital–demographic group.

Protecting population and dataset properties. Several lines of work have recognized that individual privacy does not necessarily hide sensitive properties of populations. Pufferfish and Blowfish generalize privacy definitions by allowing application-specific secrets, discriminative pairs, and public constraints (Kifer & Machanavajjhala, 2014; Pierquin et al., 2024; He et al., 2014). More directly related to our motivation, attribute privacy considers global dataset attributes such as a hospital's mortality rate (Zhang et al., 2022), while distribution privacy studies mechanisms that

conceal properties of the underlying data distribution, including demographic information (Chen & Ohrimenko, 2023). Recent work on label-distribution privacy in cross-silo federated learning similarly aims to prevent disclosure of a client's label composition (Jiang et al., 2024a).

DPAG shares the objective of protecting information above the individual-record level, but differs in its formulation. Rather than protecting a specified statistic or distributional parameter, DPAG defines adjacency through arbitrary replacement of one predefined group cell. It therefore retains the usual worst-case DP semantics while changing the atomic privacy unit to match the group whose characteristics are considered sensitive.

Differential privacy in federated learning. Federated learning allows multiple clients to train a shared model without centralizing their raw data (McMahan et al., 2017), but model updates can still reveal information about local training sets, including membership and population properties (Melis et al., 2019). Therefore, differential privacy has been applied at several granularities. Record-level approaches protect individual examples, while client- or user-level approaches clip and perturb complete client updates (McMahan et al., 2018). These choices correspond to different privacy units and need not protect an intermediate object such as a demographic subgroup within one client.

DPAG targets this intermediate granularity. Each client forms bounded group-specific contributions before aggregation, so that privacy is calibrated to a client–group cell rather than to an individual record or the complete client. This is complementary to secure aggregation, which hides individual client messages from the server but does not by itself limit what can be inferred from the final aggregate or model (Bonawitz et al., 2017). Related approaches can also reduce privacy leakage by restricting the information exchanged via federated semi-supervised learning (Bistritz et al., 2020; Abourayya et al., 2025), potentially requiring different perturbation mechanisms, such as Bernoulli noise via the XOR mechanism for hard labels (Ji et al., 2021). Thus, DPAG fits within the broader privacy-preserving federated-learning literature while protecting the distinct privacy unit of subpopulations.

3 Differentially Private Attributes of Groups (DPAG)

We introduce a mechanism that provides formal guarantees on preserving the privacy of sensitive attributes of predefined groups, such as gender, ethnicity, or pathological information across hospitals. Lifting standard differential privacy guarantees for individuals to whole groups does not scale well, leaving us with poor privacy in realistic scenarios. In the following, we propose a new approach that operates on sets of samples, here groups of individuals sharing an attribute, as opposed to individual samples. We first introduce this notion more formally for protecting a specific group in one institution, such as a hospital, in section 3.1 and extend this to protecting multiple group–institution combinations (cells) at once in section 3.2. We further show how to put this framework into practice considering different cases, including common statistical estimation (section 4.1) and machine learning tasks (sections 4.2 and 4.3).

3.1 Protection of single cells

Notation. Let individuals i have features x_i (e.g., lab results) and a group label $g_u \in \mathcal{G}$ (e.g., ethnicity). Further, let individuals belong to an institution $h_u \in \mathcal{H}$ (e.g., a hospital). Let $C = \mathcal{H} \times \mathcal{G}$ denote cells (e.g., hospital-ethnicity pairs). For each cell $c = (h, g)$, let $X_c = \{x_i : (h_i, g_i) = c\}$ denote the features of individuals belonging to cell c.

To prevent learning group characteristics (e.g., "Hospital A has mostly ethnicity g"), we protect cells as the privacy units. Intuitively, the mechanism's output should be nearly indistinguishable even if all data from one cell were different.

We say that two datasets D, D' are c-adjacent, written $D \sim_c D'$, if there exists a single cell $c \in C$ such that X_c can change arbitrarily between D and D' while all other cells are identical. This choice is made such that an adversary cannot reliably infer the characteristics (size, outcome distribution) of any specific hospital–ethnicity cell.

Definition 3.1 ((ε, δ)-Group DP)**.** *A randomized mechanism M satisfies (ε, δ)-Group DP with respect to c-adjacency if, for all measurable S,*

$$\Pr[M(D) \in S] \leq e^{\varepsilon} \Pr[M(D') \in S] + \delta \qquad \textit{for all } D \sim_c D'.$$

Thus, replacing one complete cell changes the probability of any output event by at most the multiplicative factor e^{ε} plus additive slack δ.

3.2 Protection for union of cells

Often we also want protection for unions of cells (e.g., all hospitals for one ethnicity). If M is ε, δ-Group DP at the cell level, then for a union of k cells (any k-cell change), M satisfies the standard group privacy scaling:

$$\Pr[M(D) \in S] \leq e^{k\varepsilon} \Pr[M(D') \in S] + \delta_k, \qquad \delta_k = \delta \sum_{j=0}^{k-1} e^{j\varepsilon} = \delta \cdot \frac{e^{k\varepsilon} - 1}{e^{\varepsilon} - 1}.$$

Thus, the log odds amplification scales linearly in k, and δ inflates geometrically. This gives a clear, tunable bound on the worst case inference advantage about any protected union of cells (e.g., "Does ethnicity $g^\star$ dominate across all hospitals?"). In general, if we pick ε small enough that even $k = |\mathcal{H}|$ (all hospitals for one ethnicity) $k\varepsilon$ will be small.

4 Putting DPAG to practice

Protecting whole cells naïvely as above might require very large amounts of noise. Instead, we cap each cell's contribution to every released statistic so that the cell sensitivity is small.

Bounded contribution. To ensure privacy, we compute statistics $f(D)$ as a sum of per cell contributions with hard caps and then add noise calibrated to the desired cell sensitivity. We first define a per cell count cap B, meaning at most B records from each cell may contribute to the statistic. If a cell has $n_c > B$, we uniformly subsample B of them or weight each record by B/n_c. We additionally perform per record clipping, where we clip numeric feature y to range $[a, b]$. These two steps bound each cell's total contribution and therefore yield finite cell-level sensitivity. We next instantiate this construction for several statistics and for federated learning.

Threat model: For the centralized mechanisms in Cases A and B, we assume the standard trusted-curator setting: the curator has access to the raw data and faithfully applies the prescribed mechanism, while an adversary may observe the released statistics and possess arbitrary side information. For the federated mechanism in Case C we assume honest-but-curious clients and an honest-but-curious server: all parties follow the prescribed protocol but may attempt to infer protected information from the messages and models they observe. Privacy is enforced client-side, before transmission. We do not consider malicious protocol deviations, including omitted noise, malformed updates, model poisoning, or collusion.

4.1 Case A – overall mean outcome

Let y be a feature, then we obtain an estimated mean outcome of data D subsampling B samples from each cell $c \in C$

$$\hat{\mu}(D) = \frac{1}{|C|B} \sum_{c \in C} \sum_{i \in \text{subsample}_B(X_c)} \text{clip}(y_i, [a, b]).$$

The sensitivity to a whole cell changing becomes at most

$$\frac{B(b-a)}{|C|B} = \frac{b-a}{|C|}.$$

For pure ε-DP, we can release

$$\hat{\mu}(D) + \text{Lap}\left(\frac{b-a}{|C|\varepsilon}\right).$$

For approximate DP, we can instead add Gaussian noise $\mathcal{N}(0, \sigma^2)$ with

$$\sigma = z \frac{b-a}{|C|},$$

where the noise multiplier z is calibrated by a Gaussian-mechanism privacy accountant to meet the target (ε, δ) guarantee. As the number of cells $|C| = |\mathcal{H}||\mathcal{G}|$ grows, less noise is required to protect any one cell.

4.2 Case B – institution-level means

For each institution h, we estimate the capped, cell-balanced mean as

$$\hat{\mu}_h(D) = \frac{1}{|\mathcal{G}|B} \sum_{g \in \mathcal{G}} \sum_{u \in \text{subsample}_B(X_{(h,g)})} \text{clip}(y_u, [a, b]).$$

Then, a change to one cell (h^*, g^*) changes coordinate h^* by at most $\frac{b-a}{|\mathcal{G}|}$. For the released vector $\hat{\mu} \in \mathbb{R}^{|\mathcal{H}|}$, the ℓ_2-sensitivity is therefore

$$\frac{b-a}{|\mathcal{G}|}.$$

Adding independent Laplace noise with scale $\frac{b-a}{|\mathcal{G}|\varepsilon}$ to each coordinate gives pure ε-DP. Alternatively, Gaussian noise with

$$\sigma = z \frac{b-a}{|\mathcal{G}|},$$

where z is calibrated to the desired (ε, δ) guarantee, gives approximate DP. Thus, the released vector is protected with respect to arbitrary replacement of one hospital–group cell; unions of cells follow the group-privacy scaling in Section 3.2.

In relation to standard DP, if we have individual-level DP with parameter ε_{ind}, then any change to k individuals (not necessarily a full cell) is bounded by $k\varepsilon_{\text{ind}}$, $\delta \sum_{j=0}^{k-1} e^{j\varepsilon_{\text{ind}}}$. This is weaker than cell DP when cells are large, and does not by itself prevent revealing a cell's characteristics.

4.3 Case C – Federated Learning

We modify standard FedAvg (McMahan et al., 2017) so that each client's update is a balanced average over its groups with per-group clipping following the ideas above. Here, each client adds Gaussian noise calibrated to cell sensitivity before transmitting its update. Let a model be defined by parameters $\theta \in \mathbb{R}^d$ and $t = 1 \dots T$ the rounds of FL. We provide an overview of Client- and Server-sided updates for federated learning in Figure 1.

Suppose datasets differ on a single cell (h^*, g^*). Only one group term inside one client can change. Because both the "before" and "after" per-group deltas are clipped to norm at most S, the change in that client's balanced update is at most

$$\frac{2S}{|\mathcal{G}|}$$

Client-side DPAG update at hospital h

Require: Global model θ_t, groups $\mathcal{G}$, clipping bound S, number of participating clients m_t, noise multiplier z

Ensure: Private client update $\widetilde{v}_h$

for each group $g \in \mathcal{G}$ **do**
  if group g is present at hospital h **then**
    Construct local objective using up to B samples from group g
    Initialize local model from θ_t
    Run local optimization to obtain $\Delta_{h,g}$
    $\widetilde{\Delta}_{h,g} \leftarrow \text{Clip}_{\ell_2}(\Delta_{h,g}, S)$
  else
    $\widetilde{\Delta}_{h,g} \leftarrow 0$
  end if
end for
$v_h \leftarrow \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \widetilde{\Delta}_{h,g}$
$\Delta_{\text{cell}} \leftarrow \frac{2S}{|\mathcal{G}|}$
$\xi_{h,t} \sim \mathcal{N}\left(0,\, z^2 \Delta_{\text{cell}}^2 I_d\right)$
$\widetilde{v}_h \leftarrow v_h + \xi_{h,t}$
Send $\widetilde{v}_h$ as a securely aggregated share

Server-side across hospitals

$\widetilde{v}_t \leftarrow \frac{1}{m_t} \sum_{h \in \mathcal{P}_t} \widetilde{v}_h$
$\theta_{t+1} \leftarrow \theta_t - \eta_t \widetilde{v}_t$

Figure 1: **DPAG federated learning round.** Privacy is enforced locally on group-balanced client updates; the server performs standard federated aggregation.

in ℓ_2 norm. Thus the ℓ_2 sensitivity of v_h to any one cell is

$$\Delta_{\text{cell}} = \frac{2S}{|\mathcal{G}|}.$$

Adding Gaussian noise $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma = z\Delta_{\text{cell}}$ implements the Gaussian mechanism with noise multiplier z. For a single round, this is (ε_t, δ)-DP with

$$\varepsilon_t \approx \frac{\alpha}{2z^2} + \frac{\ln(1/\delta)}{\alpha - 1}.$$

The subsequent server-side aggregation and model update are post-processing of these private client messages and therefore do not weaken the guarantee.

Instead, we can use Rényi DP (RDP) to achieve composition across T rounds. Specifically, each round contributes $\varepsilon_t^{(\alpha)} = \frac{\alpha}{2z^2}$ RDP at order α. We can then sum RDP over rounds $\sum_t \varepsilon_t^{(\alpha)}$ and convert to (ε, δ) using

$$\varepsilon = \sum_t \varepsilon_t^{(\alpha)} + \frac{\ln(1/\delta)}{\alpha - 1}$$

minimizing over α.

In the version above group balancing is essential, where we need to average over groups with a fixed weight $1/G$. This fixed factor is needed as dividing e.g. by the group's sample count would leak group sizes. Furthermore, any per-group clipping must precede client-level clipping, where the former enforces the desired cell-sensitivity and the latter would enforce client-sensitivity.

This yields a formal, implementable federated learning pipeline in which every client update is differentially private with respect to arbitrary replacement of one hospital–group cell. The aggregated updates and resulting global models inherit this guarantee by post-processing.

5 Experiments

Our experiments test the central claim of DPAG: when the sensitive object is a predefined group or hospital–group cell, directly bounding that cell's contribution can retain substantially more utility than mechanisms calibrated to a finer or stronger privacy unit. We therefore distinguish carefully between *privacy budgets* and *privacy units*. In the centralized setting, record-level DP requires group-privacy scaling to protect a complete cell. In the federated setting, client-update DP protects a stronger unit than a within-client cell and therefore already implies the corresponding cell guarantee, but pays the sensitivity cost of perturbing the entire client update. Full preprocessing, optimization, accounting, ablations, and additional experiments are deferred to Appendix A. In section 5.1 we discuss a centralized setting for group protection, in section 5.2 we study a Federated Learning scenario where hospital–group cells are the privacy protection target.

5.1 Centralized prediction: direct cell protection versus group-privacy scaling

We first isolate the core statistical mechanism in a centralized clinical prediction task using processed NHANES data (CDC, 2010). Each protected cell is an intersectional demographic group defined by sex, ethnicity, and (binned) age. Here, the training split contains $|\mathcal{C}| = 60$ non-empty cells, with largest cell size $k_{\max} = 375$. Protected attributes define privacy units only and are excluded from the predictive features. We train logistic regression and report AUROC, AUPRC, and Brier score on held-out data.

In the following, we set the target cell-level privacy as $(\epsilon, \delta) = (4, 0.01)$. We compare between (i) non-private training, (ii) standard individual-level DP-SGD, which protects one record, (iii) a cell-equivalent DP-SGD stress test, which converts the desired cell budget into an individual-level budget using classical group privacy for the largest cell, and (iv) DPAG.

In table 1 we observe that standard individual DP-SGD remains feasible at the record level (AUROC 0.860), but its guarantee does not protect an entire demographic cell. Similarly, forcing standard

Table 1: **Centralized NHANES results** at target cell-level privacy $(\epsilon, \delta) = (4, 0.01)$. Individual DP-SGD protects records, not complete cells, the cell-equivalent DP-SGD row reports the achieved group-privacy conversion. For classical group-privacy scaling (†) the achieved converted δ exceeds one and therefore does not satisfy the target cell-level guarantee.

Method	AUROC ↑	AUPRC ↑	Brier ↓	ϵ	Privacy unit
Non-private	0.917	0.606	0.062	–	none
Individual DP-SGD	0.860	0.401	0.163	3.76	individual
Cell-equiv. DP-SGD	0.616	0.185	0.270	48.3	cell†
DPAG	**0.787**	**0.422**	**0.158**	**4.00**	**cell**

DP-SGD to imply the target guarantee for all 375 members of the largest cell requires an extremely small base privacy budget. When we do a group-privacy conversion, the achieved cell-level guarantee is vacuous and utility falls to AUROC 0.616. In contrast, our approach DPAG satisfies the target cell-level budget by construction and obtains AUROC 0.787, AUPRC 0.422, and Brier score 0.158. Thus, the practical difficulty is not group protection, but obtaining it by first protecting every individual strongly enough that classical group privacy remains meaningful for a large group. DPAG avoids this dependence on $k_{\max}$ by making the cell itself the privacy unit.

5.2 Federated medical imaging prediction

We next evaluate DPAG in the cross-silo setting for which it is designed, protecting demographic subgroup *within a particular institution*. We use the UTSW-Glioma collection (Reddy et al., 2026a) with 625 patients and binary glioma segmentation from FLAIR MRI (Reddy et al., 2026b). To make repeated privacy experiments tractable while retaining a non-trivial neural segmentation task, we use one deterministic representative axial slice per patient (the slice with largest tumor-mask area) and a fixed 160×160 crop. We use a patient-level split of 435/63/127 training/validation/test cases and virtually partition the training set into ten hospitals.

Our protected demographic groups are sex crossed with five fixed age categories (< 40, 40–49, 50–59, 60–69, and ≥ 70), yielding $|\mathcal{G}| = 10$ public group slots per hospital. A privacy cell is therefore $c = (h, g)$, i.e., one hospital–demographic subgroup. Across the training partition, 91 of 100 cells are non-empty and the largest contains 33 patients. Absent groups contribute a zero update, and every client is normalized by the fixed factor $1/|\mathcal{G}|$ as required by the federated mechanism above.

We train a compact MONAI 2D U-Net and compare non-private FedAvg, DPAG, and client-update DP at RDP-calibrated $\epsilon \in \{8, 16, 32\}$ with $\delta = 0.01$. Client-update DP clips each complete hospital update and protects that stronger client-level unit; consequently, it also implies the same numerical guarantee for any change confined to one hospital–group cell. DPAG instead clips and balances group-specific updates so that the target hospital–group cell is the direct privacy unit. All reported imaging results use three independent seeds and a fixed segmentation threshold of 0.5. We calibrate all private runs using RDP; for completeness, we also report the tighter post-hoc Gaussian-DP (GDP) bound for the same Gaussian releases at $\delta = 0.01$ (Dong et al., 2022). We provide details on architecture and optimization choices in section A.4. The GDP values do not correspond to additional training runs: they evaluate the same Gaussian releases using a different accountant. In the following, $\epsilon \in \{8, 16, 32\}$ always refers to the RDP-calibrated budget used to choose the noise multiplier.

The results in Table 2 show a consistent utility gap across all three privacy budgets. FedAvg reaches 0.828 ± 0.004 Dice. DPAG obtains 0.631 ± 0.030, 0.692 ± 0.043, and 0.730 ± 0.030 at cell-level $\epsilon = 8, 16, 32$, respectively. At the same numerical budgets, client-update DP obtains only 0.022 ± 0.030, 0.062 ± 0.100, and 0.109 ± 0.182. The client-update mechanism provides a strictly stronger privacy unit, so this is not a comparison of identical adjacency relations; rather, it shows the utility cost of protecting the entire client when the intended sensitive object is a hospital–group cell. For our $|\mathcal{G}| = 10$ group slots, the sensitivity difference is exactly a factor of ten: DPAG uses $2S/|\mathcal{G}|$, whereas client-update DP uses $2S$. At the same noise multiplier z, this is also a factor of ten in the standard deviation of the Gaussian noise added to each client message. This makes explicit the cost of protecting the stronger client unit; the no-noise ablation below separates this privacy cost from the structural effect of group-wise training.

Table 2: Federated UTSW-Glioma 2D segmentation. Values are mean ± sample standard deviation over three seeds using the fixed threshold 0.5. RDP ϵ is the budget used to calibrate the Gaussian noise; GDP ϵ is the tighter post-hoc bound for the same releases at $\delta = 0.01$. Client-update DP protects an entire client's clipped update and therefore also implies the stated guarantee for any within-client hospital–group cell; DPAG protects that cell directly.

Method	Cell RDP ϵ	Cell GDP ϵ	Cell δ	Test Dice ↑
FedAvg	–	–	–	0.828 ± 0.004
DPAG	**32**	**27.26**	**0.01**	$\mathbf{0.730 \pm 0.030}$
DPAG	16	12.85	0.01	0.692 ± 0.043
DPAG	8	5.94	0.01	0.631 ± 0.030
Client-update DP	32	27.26	0.01	0.109 ± 0.182
Client-update DP	16	12.85	0.01	0.062 ± 0.100
Client-update DP	8	5.94	0.01	0.022 ± 0.030

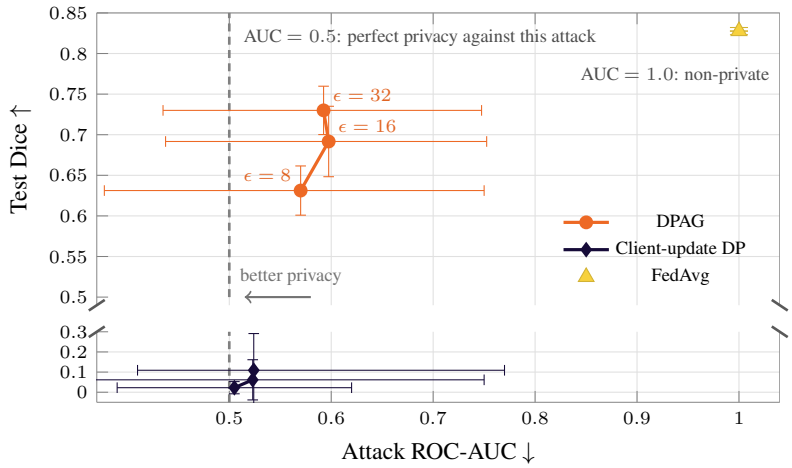

Figure 2: **Empirical privacy–utility trade-off** We consider federated UTSW-Glioma segmentation. Each point combines segmentation utility from the main three-seed experiment (vertical axis; mean with ±1 sample standard deviation) with the complete-cell replacement audit (horizontal axis; point estimate with bootstrap 95% CI). The y-axis is broken between Dice 0.3 and 0.5. Both DPAG and client-update DP are statistically compatible with the attack-specific perfect-privacy reference AUC 0.5 at all tested private budgets, but DPAG retains substantially higher utility. Non-private FedAvg is perfectly distinguishable under the same attack.

We complement the worst-case DP guarantee with a pre-specified complete-cell replacement inference audit (Figure 2; full protocol in Appendix A.6). The adversary distinguishes two disjoint candidate states of one protected hospital–group cell while all other training data remain identical. Figure 2 places the utility and attack results in a single privacy–utility plane. Non-private FedAvg is perfectly distinguishable (ROC-AUC 1.000). With Gaussian privacy noise, DPAG yields AUCs 0.570, 0.598, and 0.593 for $\epsilon = 8, 16, 32$, respectively, while client-update DP yields 0.505, 0.523, and 0.524. All six private confidence intervals include the chance-level value 0.5 and are therefore statistically compatible with perfect privacy against this fixed attack. The difference is utility: client-update DP reaches this regime with Dice between 0.022 and 0.109, whereas DPAG retains Dice between 0.631 and 0.730. This is an attack-specific empirical statement: it does not exclude the possibility that a stronger attack could achieve better-than-chance discrimination, and it does not replace the formal DP guarantee.

As a structural ablation, DPAG without Gaussian noise reaches AUC 0.708 (95% CI $[0.510, 0.872]$), showing that clipping and group-wise training alone do not account for the protection observed under private DPAG. This no-noise ablation reaches 0.764 ± 0.017 Dice, compared with 0.828 ± 0.004 for

FedAvg, indicating that part of the utility gap is caused by the group-wise training and balancing itself, while the remaining loss under private DPAG is due to Gaussian noise.

In summary, in centralized learning direct cell protection avoids the group-size penalty incurred when a record-level mechanism is lifted to a complete cell. In federated learning, client-level clipping already implies cell protection but has sensitivity $2S$; DPAG's group-specific clipping and fixed balancing reduce the target-cell sensitivity by the public factor $|\mathcal{G}|$. The empirical replacement attack gives a complementary view: both private mechanisms are statistically compatible with chance-level discrimination under the fixed attack, but only DPAG retains useful segmentation performance. Together, these results show that the relevant difference is not whether privacy can be obtained, but at which privacy unit and corresponding utility cost it is obtained.

6 Discussion and Conclusion

Discussion. We introduce DPAG to provide differential privacy for group attributes while avoiding the dependence on group size incurred when individual-level DP is lifted to groups through classical group privacy. Our experiments show that the framework can be applied in both centralized prediction and federated learning with distributed institutions. On the considered biomedical tasks, *DPAG retains substantially better utility than the relevant standard-DP alternatives*. In the centralized setting, lifting record-level DP to a complete cell through classical group privacy becomes ineffective for large cells. In the federated setting, client-update DP implies protection of a within-client cell, but protects the unit of the entire client and therefore incurs much larger sensitivity, making DPAG an attractive alternative that retains much better performance than client-update DP.

While we showed applications of our work in typical high-stakes scenarios in the biomedical domain, it would make for interesting future work to explore this framework for other types of data, e.g. containing discrete variables for which Bernoulli noise (Ji et al., 2021) serves as a candidate for our framework. DPAG assumes that the groups to be protected are specified before training. Extending the framework to more flexible or overlapping group definitions makes for interesting future work.

Limitations. Our experiments clearly show that private full-model training still remains challenging: DPAG does not recover non-private utility, and tighter privacy budgets incur a visible cost, albeit much better than existing alternatives. The important result is that matching the privacy unit to the intended sensitive object can materially improve utility. In the centralized case this avoids group-size scaling from records to cells, while in the federated case it avoids paying the sensitivity of an entire client update. Our complete-cell replacement audit is consistent with the formal protection—for all three private DPAG budgets, the fixed attack is statistically compatible with chance-level discrimination, i.e. perfect privacy against this particular attack, while it perfectly separates FedAvg—but this empirical result should not be confused with a universal privacy guarantee: failure of one attack cannot establish absence of leakage to all adversaries. Moreover, protecting finer-grained groups reduces the number of samples available for each group-specific update, potentially exacerbating the challenges of federated learning from small local datasets (Kamp et al., 2023). Similarly, differential privacy is still no golden hammer to ensure privacy. While we are able to scale to interesting scenarios such as real-world biomedical image segmentation in a federated setup, more complex related tasks such as full-volume 3D segmentation remain difficult under privacy constraints and show substantially lower utility than their non-private baselines (section A.5). Yet, also in this exploratory comparison DPAG retains substantially better utility than complete client-update perturbation.

Conclusion. We introduced DPAG, a differential privacy mechanism to protect attributes of groups in both single- and multi-institutional scenarios. Its design, aimed at overcoming poor scaling of individual-level DP when group sizes grow, allows us to give classical (ϵ, δ) guarantees on the privacy of cells (combinations of institution $\times$ group labels) and empirically shows that group privacy is feasible. **This mechanism lays the basis for new privacy guarantees for protecting groups, not just their members, in practice.**

USE OF LLMS

In this work, we used LLMs to generate snippets of text for particular sections, from which we developed the final version of text manually. We further used LLMs to generate code, giving a first implementation of our method and an outline of experimental setup. Code was then manually checked, corrected, and partially reworked.

STATEMENT ON REPRODUCIBILITY

To make our work reproducible, we provide details on Experimental Setup and Implementation in Appendix A. We will further make code publicly available after acceptance.

ETHICS STATEMENT

This work uses existing released NHANES and UTSW-Glioma datasets and does not involve new data collection. Demographic attributes are used to define the privacy units studied in this work. The complete-cell replacement attacks are conducted only as controlled evaluations of privacy leakage. We do not release additional patient-level data.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016. doi: 10.1145/2976749.2978318.

Amr Abourayya, Jens Kleesiek, Kanishka Rao, Erman Ayday, Bharat Rao, Geoffrey I. Webb, and Michael Kamp. Little is enough: Boosting privacy by sharing only hard labels in federated semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15): 15293–15301, 2025.

Sushant Agarwal, Gautam Kamath, Mahbod Majid, Argyris Mouzakis, Rose Silver, and Jonathan Ullman. Private mean estimation with person-level differential privacy. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2819–2880. SIAM, 2025.

Hilal Asi, John Duchi, and Omid Javidbakht. Element level differential privacy: The right granularity of privacy. In *AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI)*, 2022.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 394–403. PMLR, 2018.

Ilai Bistritz, Ariana Mann, and Nicholas Bambos. Distributed distillation for on-device learning. *Advances in Neural Information Processing Systems*, 33:22593–22604, 2020.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *ACM Conference on Computer and Communications Security (CCS)*, 2017. doi: 10.1145/3133956.3133982.

Robert Istvan Busa-Fekete, Andres Munoz Medina, Umar Syed, and Sergei Vassilvitskii. Label differential privacy and private training data release. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3233–3251. PMLR, 2023.

CDC. National health and nutrition examination survey data. Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, 2010.

Michelle Chen and Olga Ohrimenko. Protecting global properties of datasets with distribution privacy mechanisms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7472–7491, 2023.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, pp. 1–12. Springer, 2006. doi: 10.1007/11787006_1.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. doi: 10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14.

Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. LabelDP-Pro: Learning with label differential privacy via projections. In *International Conference on Learning Representations*, 2024.

Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *ACM SIGMOD International Conference on Management of Data*, 2014.

Tianxi Ji, Pan Li, Emre Yilmaz, Erman Ayday, Yanfang Ye, and Jinyuan Sun. Differentially private binary- and matrix-valued data query: An XOR mechanism. *Proceedings of the VLDB Endowment*, 14(5):849–862, 2021.

Yangfan Jiang, Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, and Beng Chin Ooi. Protecting label distribution in cross-silo federated learning. In *IEEE Symposium on Security and Privacy*, pp. 4828–4847, 2024a. doi: 10.1109/SP54263.2024.00113.

Yangfan Jiang, Xinjian Luo, Yuncheng Wu, Xiaokui Xiao, and Beng Chin Ooi. Protecting label distribution in cross-silo federated learning. In *IEEE Symposium on Security and Privacy (SP)*, pp. 4828–4847, 2024b. doi: 10.1109/SP54263.2024.00113.

Michael Kamp, Jonas Fischer, and Jilles Vreeken. Federated learning from small datasets. In *International Conference on Learning Representations (ICLR)*, 2023.

Fumiyuki Kato, Li Xiong, Shun Takagi, Yang Cao, and Masatoshi Yoshikawa. ULDP-FL: Federated learning with across-silo user-level differential privacy. *Proceedings of the VLDB Endowment*, 17 (11):2826–2839, 2024. doi: 10.14778/3681954.3681966.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1), 2014. doi: 10.1145/2514689.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations (ICLR)*, 2018.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *IEEE Symposium on Security and Privacy*, pp. 691–706, 2019. doi: 10.1109/SP.2019.00029.

Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017. doi: 10.1109/CSF.2017.11.

Clément Pierquin, Aurélien Bellet, Marc Tommasi, and Matthieu Boussard. Rényi pufferfish privacy: General additive noise mechanisms and privacy amplification by iteration via shift reduction lemmas. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40762–40794. PMLR, 2024.

Divya Reddy, Niloufar Saadat, James Holcomb, Benjamin Wagner, Nghi Truong, Jason Bowerman, Kimmo Hatanpaa, Toral Patel, Marco Pinho, Fang Yu, Kuan Zhang, Sadeem Lodhi, Ananth Madhuranthakam, Chandan Ganesh Bangalore Yogananda, and Joseph Maldjian. The university of texas southwestern glioma mri dataset with molecular marker characterization and segmentations (utsw-glioma), 2026a.

Divya D Reddy, Niloufar Saadat, James M Holcomb, Benjamin C Wagner, Nghi C Truong, Jason Bowerman, Kimmo J Hatanpaa, Toral R Patel, Marco C Pinho, Fang Yu, et al. The university of texas southwestern glioma dataset-mri, molecular markers and segmentations. *Scientific Data*, 2026b.

Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. Attribute privacy: Framework and mechanisms. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 757–766, 2022. doi: 10.1145/3531146.3533139.

Table 3: Federated mechanisms and privacy units. Client-update DP protects the complete clipped client update; this is a stronger adjacency relation than changing one hospital–group cell within that client and therefore implies the same (ϵ, δ) guarantee for each such cell. DPAG targets the cell directly and has lower sensitivity by the fixed group-balancing factor.

Method	Direct privacy unit	Client-message sensitivity	Cell guarantee
FedAvg	none	–	none
Client-update DP	complete client update	$2S$	implied
DPAG-FL	hospital–group cell	$2S/\lvert\mathcal{G}\rvert$	direct

A Additional Experimental Details and Results

A.1 Common protocol and privacy accounting

All private mechanisms are evaluated with respect to an explicitly stated privacy unit. For DPAG this unit is one protected cell; in the federated experiments a cell is a hospital–demographic pair $c = (h, g)$. The Gaussian mechanism is applied independently to each client's outgoing update before server-side aggregation. A record-level guarantee does not in general protect a complete multi-record cell without group-privacy conversion. By contrast, a client-level mechanism whose adjacency permits arbitrary replacement of one client's data protects a stronger unit than any change confined to a single cell within that client; hence the same client-level (ϵ, δ) guarantee also holds for such a cell change.

For a Gaussian mechanism with noise multiplier z, one release contributes RDP $\epsilon^{(\alpha)} = \alpha/(2z^2)$ at order α. Across T releases, RDP composes additively and is converted to (ϵ, δ) by minimizing

$$\epsilon = \min_{\alpha > 1} \left[T \frac{\alpha}{2z^2} + \frac{\log(1/\delta)}{\alpha - 1} \right].$$

The implementation uses the equivalent continuous-order optimum

$$\epsilon(z, T, \delta) = q + 2\sqrt{q \log(1/\delta)}, \qquad q = \frac{T}{2z^2},$$

and solves this expression for the smallest z meeting the requested target. We do not claim privacy amplification from within-cell minibatching or from the per-cell count cap. In the main UTSW-Glioma experiment all ten clients participate in every round, so no client-subsampling amplification is used either.

For the centralized record-level baseline, obtaining a guarantee for a cell of up to $k_{\max}$ records requires classical group-privacy conversion. We therefore convert the requested cell budget $(\epsilon_{\text{cell}}, \delta_{\text{cell}})$ to the base record-level target

$$\epsilon_{\text{base}} = \frac{\epsilon_{\text{cell}}}{k_{\max}}, \qquad \delta_{\text{base}} = \frac{\delta_{\text{cell}}}{\sum_{j=0}^{k_{\max}-1} \exp(j\epsilon_{\text{base}})}.$$

This conversion is not needed for the federated client-update baseline because changing one within-client cell is already contained in its stronger client-level adjacency relation.

A.2 Centralized NHANES details

We use the same processed NHANES prediction task as in the main paper. Protected cells are intersectional sex–race/ethnicity–age groups, producing 60 non-empty cells in the training split and a maximum cell size of $k_{\max} = 375$. Logistic regression is used throughout. Protected-cell identifier columns are used only to define privacy units and are excluded from the predictive feature vector.

For central DPAG, every noisy optimization step forms one contribution per protected cell. At most $B = 16$ records contribute to each cell-specific gradient, which is clipped to ℓ_2 norm $S = 1$. The clipped cell gradients are averaged with fixed weight $1/|\mathcal{C}|$, giving cell sensitivity $2S/|\mathcal{C}| = 2/60$. We run $T = 80$ noisy updates with learning rate 10^{-2}. RDP accounting is used for composition,

Table 4: Federated NHANES results at $(\epsilon, \delta) = (4, 0.01)$. Client-update DP provides a client-level guarantee and hence also protects every within-client hospital–group cell at the stated budget; DPAG-FL protects the hospital–group cell directly.

Method	AUROC ↑	AUPRC ↑	Brier ↓
FedAvg	0.780	0.365	0.186
Client-update DP	0.580	0.151	0.443
DPAG-FL	0.815	0.425	0.197

and the Gaussian noise multiplier is calibrated to target cell-level $(\epsilon, \delta) = (4, 0.01)$; the resulting multiplier is approximately $z = 8.06$.

The cell-equivalent DP-SGD baseline applies the classical group-privacy conversion to the largest cell. Its required individual-level target is $\epsilon_{\text{ind}} = 4/375$ and $\delta_{\text{ind}} = 0.01/\sum_{j=0}^{374} \exp(j\epsilon_{\text{ind}})$. In practice the calibrated mechanism does not achieve a meaningful converted cell-level guarantee: the resulting group-level ϵ is approximately 48.3 and the corresponding untruncated group-privacy δ expression exceeds one. We therefore report this row only as a feasibility stress test, not as a valid cell-private mechanism.

A.3 Federated NHANES experiment

The original federated tabular experiment is retained as an auxiliary validation of the federated mechanism. We partition the processed NHANES training data into 12 pseudo-hospitals. Protected groups are sex crossed with race/ethnicity, giving $|\mathcal{G}| = 12$ public group slots per hospital and hospital–sex–race privacy cells. The largest cell contains $k_{\max} = 540$ records. We train logistic regression for $T = 100$ rounds with $m_t = 6$ participating hospitals per round, one local epoch, client learning rate 10^{-2}, and server learning rate one.

DPAG-FL uses clipping norm $S = 0.025$. With fixed group balancing, the per-round sensitivity to one hospital–group cell is $2S/|\mathcal{G}|$, and RDP composition is calibrated to target cell-level $(4, 0.01)$, giving noise multiplier $z \approx 8.98$. The client-update DP baseline instead clips and perturbs each complete client update with sensitivity $2S$. Its client-level $(4, 0.01)$ guarantee also implies $(4, 0.01)$ for every within-client cell.

The single-run results reproduce the qualitative mechanism-level behavior: DPAG-FL remains close to the non-private utility reference, whereas perturbing complete client updates loses substantial utility. Because the newer imaging experiment provides replicated neural-network results and aligns more directly with the medical motivation, we keep this tabular FL experiment in the appendix.

A.4 UTSW-Glioma 2D segmentation details

Data and preprocessing. We use 625 UTSW-Glioma patients with FLAIR MRI and binary tumor masks, split at patient level into 435 training, 63 validation, and 127 test cases. To obtain a tractable repeated-training benchmark, each volume is converted deterministically to one representative 2D example. FLAIR intensities are clipped at the patient-volume 1st and 99th percentiles and linearly scaled to $[0, 1]$, and the tumor label is binarized. We then choose the axial slice with the largest ground-truth tumor-mask area. The crop center is the centroid of the tumor mask on that slice, and the image and mask are cropped or zero-padded to 160×160. Both slice selection and crop centering therefore use the ground-truth mask; this is a controlled supervised benchmark rather than a deployment-time preprocessing pipeline. The resulting fixed tensors are cached before federated training, so the reported 2D runs use no stochastic spatial augmentation and each patient contributes exactly one cached slice per local epoch.

The training data are partitioned into ten pseudo-hospitals. Protected demographic groups are sex crossed with five fixed age categories (< 40, 40–49, 50–59, 60–69, ≥ 70), yielding $|\mathcal{G}| = 10$ public group slots per hospital. Of the 100 possible hospital–group cells, 91 are non-empty; the median non-empty cell size is 3 and the maximum is $k_{\max} = 33$. Table 5 gives the complete training partition. Missing groups contribute a zero update, so every DPAG client update uses the same public normalization factor $1/10$.

Table 5: Training-set counts for the ten hospital–demographic cells per pseudo-hospital in the UTSW-Glioma 2D experiment. Columns are sex crossed with the five fixed age categories. Zero entries correspond to public group slots that contribute a zero update under DPAG-FL.

	Female					Male					
Hospital	40s	50s	60s	≥ 70	< 40	40s	50s	60s	≥ 70	< 40	Total
0	2	2	5	4	6	8	9	4	1	6	47
1	3	4	4	5	9	3	5	9	2	15	59
2	2	0	2	0	3	0	6	9	4	1	27
3	1	6	19	14	2	2	2	3	4	1	54
4	4	1	2	0	1	2	2	6	2	1	21
5	1	3	4	9	0	1	6	4	2	2	32
6	2	3	1	0	2	1	5	2	4	3	23
7	2	1	3	3	7	7	4	4	1	6	38
8	4	6	3	6	8	6	3	1	0	5	42
9	2	0	0	1	5	5	20	33	23	3	92
Total	23	26	43	42	43	35	62	75	43	43	435

Model and optimization. We use a compact MONAI 2D U-Net with channels $(2, 4, 8, 16)$ and one residual unit, containing 3,352 trainable parameters. The loss is MONAI `DiceCELoss` with sigmoid output, squared prediction in the Dice term, and equal Dice and cross-entropy weights. Every local trajectory is optimized with AdamW, learning rate 5×10^{-4} and weight decay 10^{-5}. Unless stated otherwise, all ten clients participate in each of $T = 10$ rounds; local batch size is four, each local trajectory runs for 48 epochs, the evaluation batch size is 32, and the server applies the aggregated model delta with learning rate one. Mixed precision is enabled. The reported model is always the final-round model rather than a validation-selected checkpoint.

Federated mechanisms. FedAvg performs ordinary pooled local training on each participating hospital and combines client deltas using sample-size-weighted averaging. For DPAG-FL, each non-empty hospital–group cell starts an independent local model trajectory from the same current global model θ_t. Training on that cell produces a model delta $\Delta_{h,g}$, which is clipped to ℓ_2 norm $S = 1$ before any cross-group aggregation. Missing group slots contribute the zero vector. The hospital update is therefore

$$v_h = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathrm{Clip}_S(\Delta_{h,g}),$$

and participating hospitals are averaged uniformly at the server. Since the main experiment uses $B = 100 > k_{\max} = 33$, the per-cell count cap does not subsample any training cell.

The client-update DP baseline uses ordinary pooled local training rather than group-specific trajectories. Each complete client model delta is clipped to $S = 1$, and Gaussian noise calibrated to sensitivity $2S$ is added before transmission; participating clients are then averaged uniformly. This mechanism is private with respect to arbitrary replacement of one client's data. Since changing one hospital–group cell while holding the remainder of that client fixed is a special case of such a replacement, the client-level guarantee also implies the same (ϵ, δ) guarantee for the target cell.

For DPAG-FL, changing one hospital–group cell can change only one clipped term in one hospital update, giving client-message sensitivity

$$\Delta_{\mathrm{cell}} = \frac{2S}{|G|} = \frac{2S}{10} = \frac{S}{5}.$$

Gaussian noise calibrated to this sensitivity is added independently to each client update before aggregation.

The main paper reports replicated RDP-calibrated $\epsilon \in \{8, 16, 32\}$ experiments for both DPAG-FL and client-update DP, all with $\delta = 0.01$. For the three budgets, the calibrated noise multipliers are approximately 1.592, 0.934, and 0.573, respectively; because both mechanisms use the same

Table 6: Per-seed UTSW-Glioma 2D test Dice. Client-update DP and DPAG are compared at matched numerical (ϵ, δ); the former provides the stronger client-level guarantee. The no-noise DPAG row is a non-private structural ablation.

Method	Cell ϵ	Seed 0	Seed 1	Seed 2
FedAvg	–	0.823	0.828	0.832
DPAG	8	0.603	0.664	0.627
Client-update DP	8	0.008	0.002	0.057
DPAG	16	0.676	0.741	0.658
Client-update DP	16	0.008	0.000	0.177
DPAG	32	0.763	0.705	0.722
Client-update DP	32	0.008	0.000	0.319
DPAG no-noise	–	0.783	0.754	0.754

Gaussian accountant, the utility difference comes from the sensitivity and aggregation structure rather than a different numerical privacy budget.

For completeness, a μ-GDP mechanism implies the privacy profile

$$\delta(\epsilon) = \Phi\left(-\frac{\epsilon}{\mu} + \frac{\mu}{2}\right) - e^{\epsilon}\Phi\left(-\frac{\epsilon}{\mu} - \frac{\mu}{2}\right),$$

where Φ is the standard normal CDF. We numerically invert this expression at $\delta = 0.01$ to obtain the GDP ϵ values reported in the main table.

Evaluation and replication. Sigmoid probabilities are thresholded at the fixed value 0.5 to obtain binary tumor masks, and utility is the mean patient-level Dice score on the held-out test set. No validation-selected threshold is used. For every random seed, all compared methods use the same model initialization seed. The main estimates use seeds 0, 1, and 2. Reported "$\pm$" values and error bars denote the sample standard deviation across the three independently trained models.

Across the three seeds, DPAG obtains 0.631 ± 0.030, 0.692 ± 0.043, and 0.730 ± 0.030 Dice at $\epsilon = 8, 16, 32$, whereas client-update DP obtains 0.022 ± 0.030, 0.062 ± 0.100, and 0.109 ± 0.182. The non-private DPAG structural ablation, which keeps group-specific trajectories, clipping, and fixed group balancing but adds no Gaussian noise, reaches 0.764 ± 0.017 Dice. Thus some gap to FedAvg (0.828 ± 0.004) is due to the group-balanced/clipped training structure itself, while the remaining degradation under DPAG reflects privacy noise.

A.5 Exploratory full-volume 3D segmentation

We retain the earlier full-volume experiment as a stress test for high-dimensional private model updates. The same UTSW-Glioma cohort is represented by deterministic tumor-centered 96^3 image–mask tensors and partitioned into ten pseudo-hospitals. Protected groups are sex crossed with a binary age category (≤ 50 vs. > 50), yielding four group slots per hospital and maximum cell size 75. At most $B = 16$ patients contribute to a local group update. The reported private rows target $(\epsilon, \delta) = (4, 0.01)$ over $T = 10$ releases. DPAG uses the hospital–group cell as its direct privacy unit, while client-update DP protects the stronger complete-client unit and therefore also implies this cell guarantee.

The 3D results are substantially weaker than the 2D replicated experiment and are all single-seed. Client-update DP is poor, and DPAG retains more utility but remains far below FedAvg. We therefore treat these runs as evidence about the difficulty of full-model perturbation rather than as the primary quantitative result.

Tumor-centered preprocessing and full participation improve utility in these exploratory runs. At fixed total privacy budget, increasing the number of communication rounds does not necessarily help because composition requires more noise per release. Increasing local epochs can reduce the number of releases for a fixed amount of local computation, but excessive local optimization produces substantial drift. For the 3D U-Net, reducing the clipping norm from $S = 1$ to $S = 0.3$ lowers the Gaussian noise scale and improves the observed single-seed Dice to 0.212, at the cost of clipping a

Table 7: Exploratory 3D UTSW-Glioma segmentation results from the earlier full-volume study. All values are single-seed and are retained as an auxiliary stress test rather than a main-paper comparison.

Model	Method	S	Batch	Test Dice
UNETR	FedAvg	–	16	0.605
UNETR	Client-update DP	1.0	16	0.141
UNETR	DPAG-FL	1.0	16	0.177
3D U-Net	FedAvg	–	16	0.532
3D U-Net	DPAG-FL	1.0	16	0.170
3D U-Net	DPAG-FL	0.3	4	0.212

Table 8: Exploratory 3D DPAG-FL ablations. T is the number of rounds, m participating clients, E local epochs, B the per-cell cap, and S the clipping norm.

Model	Crop	T	m	E	B	S	Batch	Dice
UNETR	fixed	20	5	1	16	1.0	16	.096
UNETR	fixed	20	5	1	16	0.5	16	.091
UNETR	tumor	20	10	1	16	1.0	16	.135
UNETR	tumor	10	10	2	16	1.0	16	.177
UNETR	tumor	5	10	4	16	1.0	16	.112
UNETR	tumor	10	10	2	32	1.0	16	.163
UNETR	tumor	10	10	3	16	1.0	16	.057
UNETR	tumor	15	10	2	16	1.0	16	.114
3D U-Net	tumor	10	10	2	16	1.0	16	.170
3D U-Net	tumor	10	10	2	16	0.3	4	.212

large fraction of cell updates. These ablations motivated the more controlled 2D study used in the main paper.

A.6 Complete-cell replacement inference audit

As a mechanism-level privacy audit, we construct a strong cell-level analogue of membership inference. The target is one hospital–demographic cell (client 9, male patients aged 60–69). From its 33 original patients we form two disjoint candidate states A and B, each containing 16 patients. The two candidates are matched in tumor-burden distribution (mean foreground areas 2745 and 2687 pixels, respectively), so the hidden bit is primarily which complete candidate cell was used rather than a coarse burden shift. Every other training row, and all validation and test data, are identical between the two worlds.

For each mechanism we train 20 matched repetitions in world A and 20 in world B. The attacker receives the released model and both candidate sets and uses the pre-specified direct score $\overline{L}(A) - \overline{L}(B)$, i.e., the difference in mean segmentation loss on the two candidates. This is deliberately stronger than ordinary individual membership inference: it asks which of two complete neighboring cell states generated the released model. We report ROC-AUC across the 40 released models and bootstrap 95% confidence intervals obtained by resampling matched repeat IDs.

Table 9: Complete-cell replacement inference audit on the UTSW 2D task. For each configuration we train 20 matched repetitions in each of two neighboring worlds (40 released models total). The primary attack ranks models using the pre-specified difference between mean losses on the two disjoint candidate cell states. AUC confidence intervals are bootstrap 95% intervals over matched repeat IDs. AUC $= 0.5$ denotes random ranking for this attack; it is an empirical reference point, not a formal privacy guarantee.

Method	Models	Mean test Dice	Attack AUC (95% CI)
DPAG ($\epsilon = 8$)	40	0.596 ± 0.157	0.570 $[0.377, 0.750]$
DPAG ($\epsilon = 16$)	40	0.679 ± 0.070	0.598 $[0.438, 0.753]$
DPAG ($\epsilon = 32$)	40	0.717 ± 0.046	0.593 $[0.435, 0.748]$
DPAG clip-only ($z = 0$)	40	0.749 ± 0.030	0.708 $[0.510, 0.872]$
Client-update DP ($\epsilon = 8$)	40	0.022 ± 0.030	0.505 [0.390, 0.620]
Client-update DP ($\epsilon = 16$)	40	0.062 ± 0.100	0.523 [0.310, 0.750]
Client-update DP ($\epsilon = 32$)	40	0.109 ± 0.182	0.524 [0.410, 0.770]
FedAvg	40	0.835 ± 0.007	1.000 $[1.000, 1.000]$

FedAvg is perfectly distinguishable (AUC 1.0), demonstrating that the released non-private model retains enough information to identify the complete target-cell state in every tested repetition. Removing Gaussian noise while retaining DPAG's group-wise training and clipping reduces the attack to AUC 0.708, but its confidence interval remains just above the random-ranking reference. Under private DPAG, the observed AUCs lie between 0.570 and 0.598, and all three 95% intervals contain 0.5. We interpret 0.5 as perfect privacy against this fixed empirical attack: the attack cannot distinguish the two neighboring worlds better than chance. This remains attack-specific—it neither validates nor estimates the formal ϵ, and it does not rule out stronger attacks. Conversely, the audit is still useful as a sanity check because the same protocol cleanly detects the non-private FedAvg leakage.

A.7 Implementation and reproducibility

The segmentation experiments are implemented in Python using PyTorch and MONAI. Random seeds are applied to NumPy, PyTorch, and CUDA; for a fixed seed, compared methods start from the same model initialization. The 2D experiment uses precomputed fixed tensors, so the local training loop only shuffles cached patient examples and introduces no additional stochastic spatial augmentation. Mixed-precision training is used for the reported neural-network experiments. All reported neural-network experiments were run on a single NVIDIA H100 80 GB GPU using CUDA 12.8, Python 3.11.15, PyTorch 2.13.0, and MONAI 1.6.0.