When More Data Is Less Trustworthy: Site-Signal Alignment Failure Modes in Federated Computational Pathology

Matthew Vaishnav
Independent Researcher

June 3, 2026

Abstract

I study a failure mode in federated computational pathology: FedAvg gives more aggregation authority to clients with more samples, but sample count is not necessarily the same as task-specific site-signal alignment. Pathology labels may reflect local grading thresholds, staining and scanning workflows, case mix, annotation practices, pathologist disagreement, or historical reporting policy. A high-volume site can therefore contribute many samples while also producing a training signal that is less aligned with the declared validation objective.

I test this failure mode using simulated federations over PANDA-derived Phikon feature representations for prostate cancer grading. Validation labels are kept clean while the largest simulated client's training signal is perturbed through dominant-site label corruption and systematic ordinal threshold shift. The central question is whether sample-size-dominant aggregation becomes less safe when the dominant client's training signal is misaligned, and whether dominance-aware switching can reduce that risk.

The strongest result is a fixed detector-switch rule calibrated under dominant-site label-noise stress and evaluated on conservative ordinal threshold-shift stress. The rule keeps clean-regime switching low at 13.3% and produces statistically positive improvements at 35% and 45% conservative shift across global quadratic weighted kappa (QWK), macro-F1, and worst-site QWK. Diagnostic analysis shows that detector triggers are mainly driven by ordinal-error increase and QWK degradation, not by a single site-spread heuristic. Leave-one-diagnostic-family-out ablation shows that removing the most frequent diagnostic, mean absolute ordinal error, reduces trigger rate but does not collapse the positive 35% / 45% transfer result. Calibration-sensitivity analysis finds that 29 of 36 nearby detector settings preserve low clean-regime switching and positive 35% / 45% gains across global QWK, macro-F1, and worst-site QWK.

These results do not establish clinical readiness or real hospital deployment performance. They support a narrower claim: in simulated federated pathology experiments over real pathology-derived features, raw sample count is not equivalent to task-specific site-signal alignment, and sample-size dominance should be treated as an auditable modeling assumption rather than an automatic guarantee of aggregation safety.

1 Introduction

Federated learning is attractive in medicine because raw patient data is difficult, risky, and often impossible to centralize. In pathology, the motivation is especially clear: whole-slide images and derived feature representations may be governed by institutional policy, patient privacy requirements, storage constraints, and clinical governance. Federated learning offers a way for multiple sites to train collaboratively while keeping raw local data inside each environment.

However, federated learning does not remove the question of influence. It changes the question from "who owns the data?" to "how much should each site shape the shared model?" Standard FedAvg answers that question with a simple rule: clients with more samples receive more aggregation influence [3].

That rule is not obviously safe in computational pathology. More samples can coexist with a training signal that is less aligned with the declared validation objective. A large site may differ because of grading thresholds, staining or scanning workflow, case mix, annotation source, label workflow, local reporting practice, or patient population. These differences are not evidence that an institution is worse or less trustworthy. They are evidence that sample volume and task-specific training-signal alignment are different quantities.

I study a controlled version of that problem. I simulate multi-site federated learning over PANDA-derived Phikon feature vectors. I perturb the largest simulated site while keeping validation labels clean. The goal is to test whether FedAvg becomes vulnerable when its sample-size assumption is violated, and whether a detector-switch mechanism can identify when sample-size dominance has become unsafe.

Working thesis

In federated computational pathology, raw sample count is not the same as task-specific site-signal alignment. FedAvg can become less safe when the largest simulated pathology client has a training-label process that is misaligned with the validation objective, and dominance-aware aggregation or switching can reduce that risk under controlled stress.

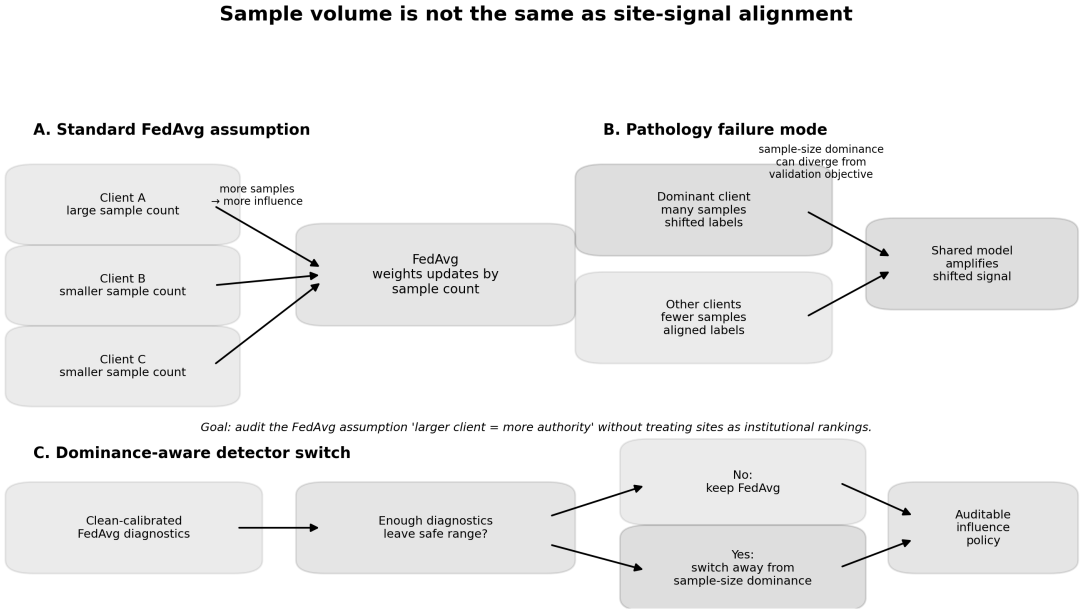

Figure 1: Problem schematic. FedAvg uses sample count as aggregation authority, but a high-volume site can have a training-label process that is less aligned with the declared validation objective.

2 Contributions

I make five main contributions.

1. **A sample-volume / site-signal alignment failure-mode framing.** I reframe dominant-client failure not as institutional reliability ranking, but as an audit of a modeling assumption that FedAvg already makes silently: larger client equals more influence.
2. **A simulated federated pathology stress test over real pathology-derived features.** I use PANDA-derived Phikon feature vectors and perturb the dominant simulated site's training signal while keeping validation labels clean.
3. **Two stress modes.** I examine dominant-site label corruption and systematic ordinal threshold shift. The latter is intended to approximate a more pathology-plausible failure mode than purely random label corruption.
4. **A fixed detector-switch transfer result.** I show that a detector rule calibrated under label-noise stress transfers to conservative ordinal threshold shift, with low clean-regime switching and positive 35% / 45% shift gains.
5. **Diagnostic and calibration robustness checks.** I support the detector result with diagnostic-frequency analysis, leave-one-diagnostic-family-out ablation, and calibration-sensitivity analysis over nearby detector settings.

3 Ethical framing: site-signal alignment, not institutional worth

This work does *not* claim that some hospitals, pathologists, or institutions are inherently more reliable, competent, or trustworthy than others.

The term *site-signal alignment* is used in a narrow modeling sense: whether a simulated client's training signal appears aligned with the declared validation objective under a specific experimental setup. A client may appear misaligned for many non-blameworthy reasons, including different grading thresholds, staining or scanning protocol differences, local case mix, patient-population differences, annotation workflow differences, label-source differences, historical reporting practice, local clinical policy, or pathologist disagreement.

Dominance-aware aggregation should therefore not be interpreted as an institutional ranking system. It is an audit mechanism for a modeling assumption that FedAvg already makes silently: larger client equals more influence.

The ethical purpose of this work is to make that assumption visible, stress-testable, and contestable. Any real deployment would require governance, local clinical review, pathologist input, bias auditing, institutional agreement on validation objectives, prospective validation, security review, and regulatory review.

4 Experimental setup

I use PANDA-derived Phikon feature representations and simulate multi-site federated learning over pathology-derived feature vectors. PANDA is a prostate cancer grading dataset [1]. Phikon is a pathology foundation model used here as the feature extractor [2].

Table 1: Experimental setup.

Category	Value
Dataset	PANDA prostate cancer grading
Feature extractor	Phikon
Readable feature files	10,611
Feature dimension	768
Task	ISUP grade prediction, classes 0–5
Federation	Simulated multi-site setting
Perturbed client	Largest simulated client
Validation labels	Kept clean during stress experiments
Metrics	Global QWK, worst-site QWK, mean-site QWK, macro-F1, accuracy
Seeds	15-seed stress studies

Quadratic weighted kappa (QWK) is used because ISUP prostate grading is ordinal. Confusing adjacent grades is less severe than confusing grade 0 with grade 5, and QWK better reflects that ordinal structure than plain accuracy alone.

The experimental design keeps validation labels clean. Perturbations are applied to the largest simulated site's training labels. This isolates the question of whether aggregation remains safe when a high-volume client's training signal becomes less aligned with the target validation objective.

5 Stress modes

5.1 Dominant-site label corruption

The first stress mode corrupts labels at the largest simulated site. This creates a controlled failure mode where the site with the most samples remains influential under FedAvg even though its training labels become less aligned with the validation objective.

The observed pattern is conditional: FedAvg remains strong in the clean setting and should not be automatically replaced; cross-site blending improves robustness under corrupted dominant-site stress; and a clean-calibrated detector can switch away from FedAvg in unsafe regimes.

5.2 Systematic ordinal threshold shift

Random corruption is useful for exposing a mechanism, but pathology disagreement is often systematic rather than random. The second stress mode applies ordinal threshold shift to the dominant site's training labels. Aggressive shift moves selected labels upward by one ISUP grade when possible. Conservative shift moves selected labels downward by one ISUP grade when possible.

The conservative threshold-shift result is the strongest transfer setting. I treat it as the headline systematic-bias result. Aggressive threshold shift is weaker and is presented as a supplementary asymmetric-stress result rather than the main claim.

6 Aggregation and detector-switch logic

FedAvg encodes a simple statistical assumption: more samples equals more authority. The detector-switch approach treats that as an assumption to audit rather than an axiom. The detector follows this logic:

1. Calibrate normal FedAvg validation behavior on clean runs.
2. Monitor validation diagnostics.
3. If enough diagnostics leave the clean-calibrated safe range, switch away from sample-size dominance.
4. Otherwise, keep FedAvg.

The fixed detector rule evaluated in the conservative threshold-shift transfer experiment was:

Setting	Value
low_quantile	0.10
high_quantile	0.80
min_trigger_count	3
use_entropy	false

The detector diagnostics include global QWK, worst-site QWK, site-QWK spread, mean absolute ordinal error, and severe ordinal error rate. Prediction entropy is available but was not used in the fixed headline rule.

7 Mathematical notation

Let there be K simulated clients. Client k has n_k local training samples and produces a local model update $\Delta\theta_k$. Standard FedAvg assigns aggregation weight

$$w_k = \frac{n_k}{\sum_{j=1}^{K} n_j} \tag{1}$$

and forms the global update

$$\Delta\theta = \sum_{k=1}^{K} w_k \Delta\theta_k. \tag{2}$$

The concern studied here is that n_k is a volume term, not an alignment term. A high-volume client may have a training-label process that is less aligned with the declared validation objective.

Let A_k denote the task-specific alignment of client k's training signal with the declared validation objective. The claim is not that A_k is an institutional property. It is a run-specific modeling quantity. The failure mode appears when the dominant sample-volume client has high n_k but reduced A_k.

Let $d_1, \ldots, d_m$ be clean-calibrated validation diagnostics. Each diagnostic has a clean safe range defined by lower and/or upper quantiles. For a run r, define

$$I_i(r) = \begin{cases} 1, & \text{if diagnostic } d_i \text{ leaves its clean-calibrated safe range,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The fixed detector rule triggers when

$$\sum_i I_i(r) \geq 3. \tag{4}$$

Let $M_{\text{clean}}(r)$ be the metric obtained by staying with the clean strategy, and $M_{\text{switch}}(r)$ be the metric obtained by the detector-switch policy. The reported delta is

$$\Delta M(r) = M_{\text{switch}}(r) - M_{\text{clean}}(r). \tag{5}$$

Positive ΔM means the detector-switch policy improved over staying with the clean strategy for that metric.

8 Fixed detector transfer result

I evaluated a fixed label-noise-calibrated detector rule on conservative threshold-shift stress. Table 2 summarizes the result.

Table 2: Fixed detector transfer to conservative ordinal threshold shift. Deltas are detector-switch performance minus the clean-strategy baseline.

Shift	Trigger rate	Global QWK delta	95% CI	Macro-F1 delta	95% CI	Worst-site QWK delta	95% CI
0%	13.3%	-0.00025	[-0.00113, 0.00062]	+0.00009	[-0.00039, 0.00057]	+0.00151	[-0.00162, 0.00463]
25%	33.3%	+0.00129	[-0.00175, 0.00432]	+0.00290	[-0.00184, 0.00765]	+0.00353	[-0.00428, 0.01133]
35%	60.0%	+0.00542	[0.00062, 0.01022]	+0.00838	[0.00272, 0.01405]	+0.00991	[0.00169, 0.01813]
45%	73.3%	+0.01053	[0.00239, 0.01866]	+0.01512	[0.00819, 0.02204]	+0.01290	[0.00547, 0.02034]

The clean 0% conservative-shift regime has low switching at 13.3%, with near-zero global QWK cost and a confidence interval crossing zero. At 25% shift, the result is directionally positive but not statistically clean because the confidence intervals cross zero. The 35% and 45% regimes are the headline positive transfer results: both show statistically positive improvements across global QWK, macro-F1, and worst-site QWK.

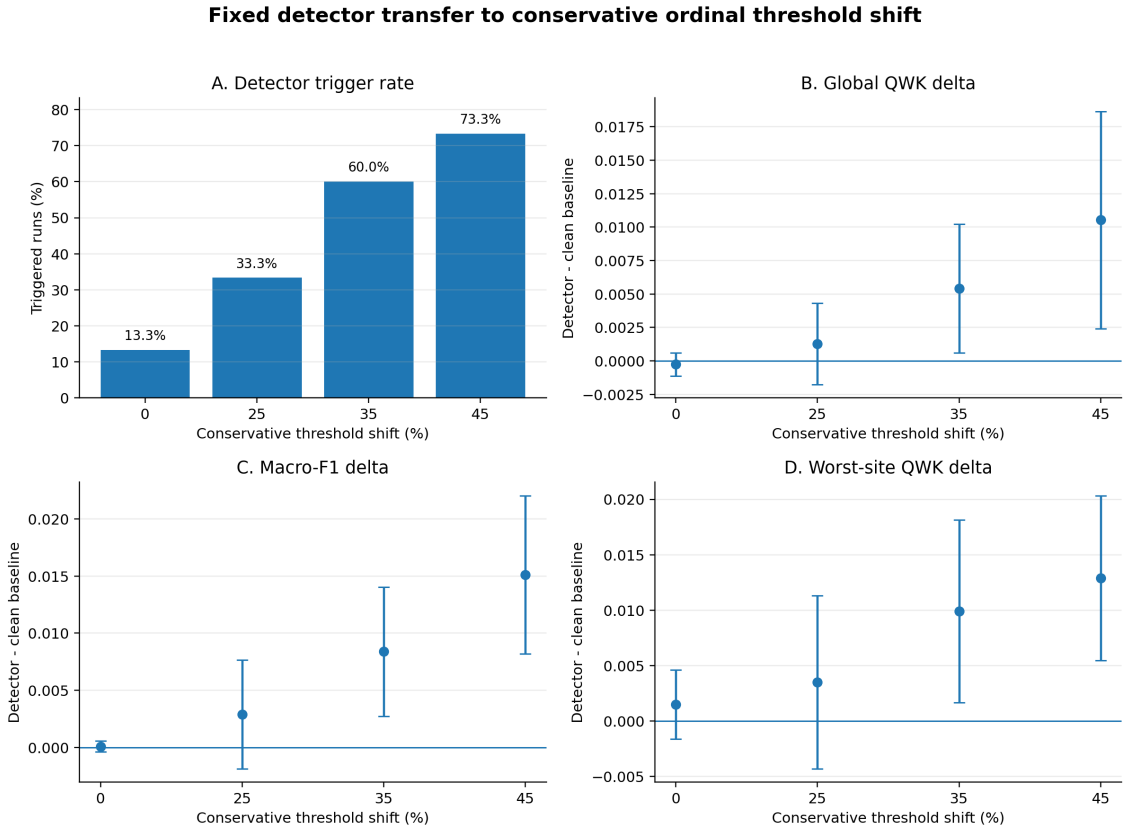

Figure 2: Fixed detector transfer to conservative ordinal threshold shift.

9 Detector diagnostic analysis

The next question is whether the detector is interpretable. If it only triggers because of one arbitrary metric, the result is less convincing. The diagnostic-frequency analysis summarizes which clean-calibrated diagnostics were crossed by the fixed detector.

Table 3: Diagnostic-frequency analysis for conservative threshold shift.

Diagnostic	Count
mean_abs_error_high	44
worst_site_qwk_low	31
global_qwk_low	27
severe_error_rate_high	22
site_qwk_spread_high	12

The detector is driven primarily by ordinal-error and QWK degradation signals, not by site-spread alone. This is mechanistically sensible in a conservative ordinal threshold-shift setting.

10 Leave-one-diagnostic-family-out ablation

The most frequent diagnostic is `mean_abs_error_high`. The key ablation question is whether the fixed-rule transfer result collapses if this diagnostic is removed. Table 4 summarizes the main comparison over the 35% and 45% conservative threshold-shift regimes.

Table 4: Leave-one-diagnostic-family-out and single-family ablation over 35% and 45% conservative threshold shift.

Variant	Mean trigger rate	Global QWK delta	Macro-F1 delta	Worst-site QWK delta	Significant global-QWK regimes	Positive global-QWK regimes
only_mean_abs_error_high	96.7%	+0.00865	+0.01462	+0.01054	2	2
full	66.7%	+0.00797	+0.01175	+0.01141	2	2
minus_site_qwk_spread_high	60.0%	+0.00797	+0.01124	+0.01092	2	2
only_global_qwk_low	60.0%	+0.00797	+0.01124	+0.01092	2	2
minus_worst_site_qwk_low	53.3%	+0.00770	+0.01011	+0.00940	2	2
only_severe_error_rate_high	50.0%	+0.00757	+0.00958	+0.00847	2	2
minus_severe_error_rate_high	63.3%	+0.00728	+0.01167	+0.01057	1	2
only_worst_site_qwk_low	76.7%	+0.00709	+0.01357	+0.00974	1	2
minus_global_qwk_low	56.7%	+0.00701	+0.01054	+0.00904	2	2
minus_mean_abs_error_high	50.0%	+0.00701	+0.01003	+0.00856	1	2
only_site_qwk_spread_high	23.3%	+0.00128	+0.00234	+0.00228	0	2

Removing `mean_abs_error_high` reduces trigger rate but does not collapse the transfer result. The signal is distributed across multiple diagnostics, especially QWK degradation and severe ordinal-error signals.

11 Calibration-sensitivity analysis

The fixed detector should not be treated as convincing if it only works at one hand-picked threshold setting. To test this, I swept nearby detector settings:

$$\begin{aligned} \texttt{low\_quantile} &\in \{0.05, 0.10, 0.15\}, \\ \texttt{high\_quantile} &\in \{0.75, 0.80, 0.85, 0.90\}, \\ \texttt{min\_trigger\_count} &\in \{2, 3, 4\}. \end{aligned}$$

This produced 36 detector configurations. A configuration was counted as robust-positive if it preserved clean trigger rate at or below 20% and positive global-QWK, macro-F1, and worst-site-QWK deltas at both 35% and 45% conservative shift. In total, 29 of 36 configurations were robust positive.

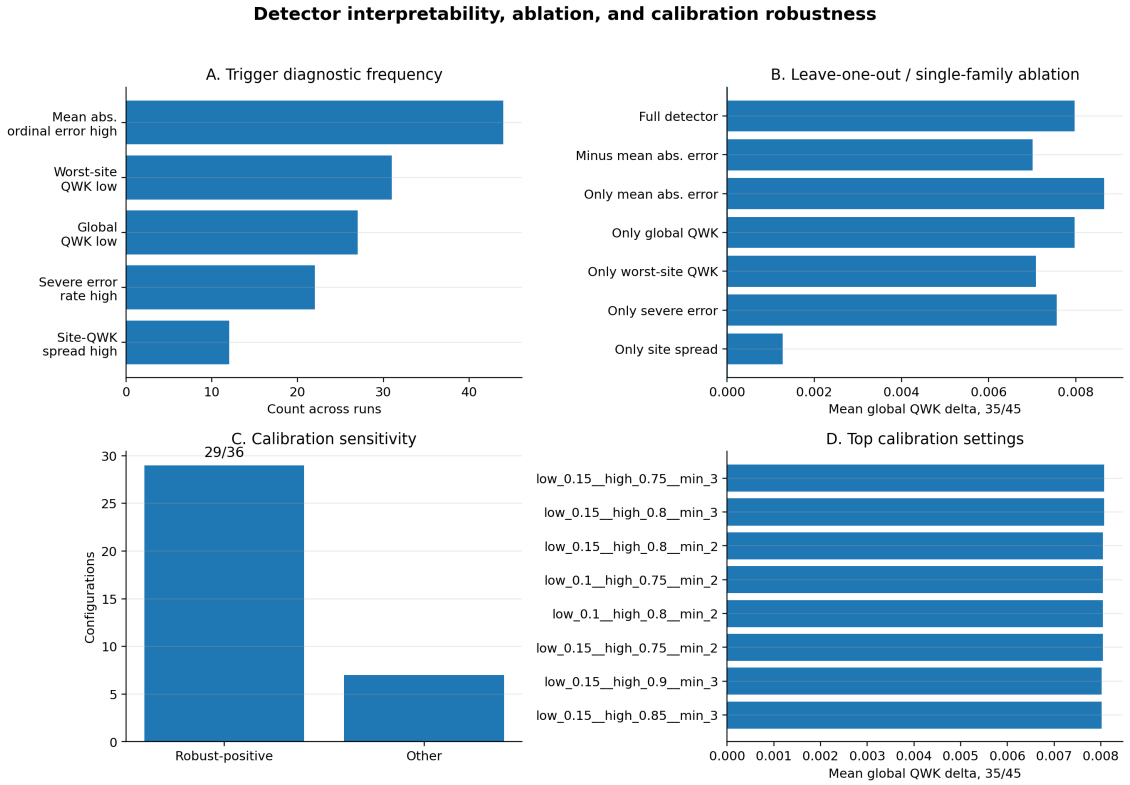

Figure 3: Detector interpretability, ablation, and calibration robustness. The transfer result is not a one-diagnostic or one-threshold artifact in the conservative threshold-shift setting.

12 Supplementary asymmetric-stress result: aggressive threshold shift

Aggressive threshold shift is a weaker and more mixed transfer setting. Under aggressive shift, the fixed detector kept the same 13.3% clean-regime trigger rate, but trigger rates remained low rather than scaling strongly with stress: 26.7% at 25%, 26.7% at 35%, and 20.0% at 45%. Global QWK and worst-site QWK were mildly positive, while macro-F1 was negative at 25% and 35% and positive only at 45%.

This asymmetry is important: the detector should not be presented as a universal stress detector. The conservative threshold-shift result is the headline; aggressive threshold shift is a supplementary weak-transfer result that shows where the detector is less responsive.

13 Relation to broader repository evidence

13.1 PCam

The repository includes PCam full-dataset validation with 85.26% test accuracy and 0.9394 AUC on the complete 32,768-sample test set. This validates the patch-level image pipeline on real pathology patches but does not validate whole-slide aggregation or hospital deployment.

13.2 PANDA and TransnnMIL

The PANDA slide-level work uses 10,611 readable Phikon feature bags with feature dimension 768. Mean-pooled Phikon plus MLP reached QWK 0.7274, gated AttentionMIL reached QWK 0.8100, and tuned TransnnMIL repeated-seed runs reached QWK 0.8155, 0.8225, and 0.8086. A stabilized warmup-cosine / gradient-clipping TransnnMIL grid produced mean best validation QWK between 0.8117 and 0.8257 across 18 runs. The supported claim is competitive behavior, not architecture superiority.

13.3 PathologyFL

PathologyFL provides the federated execution layer for simulated-site computational pathology experiments. It includes coordinator/client workflows, weighted aggregation, privacy hooks, and robustness-oriented diagnostics. These systems support research experiments but do not establish real multi-center deployment readiness.

13.4 FAIR-WEIGHTS-H

FAIR-WEIGHTS-H is an auditable institutional weighting protocol that separates training weight, validation representation priority, and monitoring priority. Its intended role is to enforce pre-specified representation and subgroup-performance constraints rather than to claim that a learned scalar weight proves institutional fairness. Current empirical status is conservative: FAIR-WEIGHTS-H runs and produces distinct weight trajectories under heterogeneous simulated sites, but it has not yet demonstrated a consistent performance or fairness advantage over simpler baselines.

14 Discussion

The main scientific point is not that cross-site blending is always superior to FedAvg. The main point is conditional: FedAvg can become less safe when the dominant client's training signal becomes misaligned with the validation objective, because FedAvg continues to assign that client high influence based on sample count.

The detector-switch result suggests one possible audit mechanism. Rather than replacing FedAvg everywhere, the detector preserves FedAvg in clean regimes and switches away when enough clean-calibrated diagnostics indicate degraded alignment. The clean-regime switch rate remains low in the conservative threshold-shift transfer result, while 35% and 45% shift regimes show positive gains.

The diagnostic analyses make the detector more credible. It is mainly driven by ordinal-error and QWK degradation signals, which are mechanistically connected to ordinal threshold shift. It does not collapse when the most frequent diagnostic is removed. It also remains stable across a neighborhood of calibration settings.

15 Limitations

The limitations are substantial. First, this is a simulated federation over pathology-derived feature vectors, not a real multi-hospital deployment. Site identity is simulated, and the perturbations are controlled. The result may not hold unchanged under naturally occurring site distributions.

Second, the data are derived from PANDA prostate cancer grading. A stronger claim would require external validation on real multi-center pathology data, such as Camelyon17 or another dataset with natural center identity.

Third, the detector uses validation diagnostics. Any real deployment would need to define how validation data are governed, where diagnostics are computed, and whether those diagnostics are allowed to leave institutional environments.

Fourth, the conservative threshold-shift result is stronger than the aggressive threshold-shift result. I headline conservative shift and treat aggressive shift as weaker or supplementary.

Fifth, the detector is not universally calibrated. The calibration-sensitivity analysis shows robustness in a local neighborhood of settings for this experiment. It does not prove universal detector calibration across datasets, diseases, institutions, scanners, or label policies.

Sixth, the experiments do not establish clinical readiness, diagnostic safety, regulatory compliance, or deployment suitability.

16 Claim boundaries

Supported claims include: FedAvg has a sample-volume / site-signal alignment failure mode in these simulated federated pathology experiments; the failure mode appears when the largest simulated client's training signal is made less aligned with the validation objective while validation labels remain clean; cross-site blending improves robustness under dominant-site

label corruption and conservative ordinal threshold shift; a fixed label-noise-calibrated detector transfers to conservative threshold-shift stress with low clean-regime switching and positive 35% / 45% shift gains; the detector is interpretable; removing `mean_abs_error_high` does not collapse the 35% / 45% transfer result; and calibration-sensitivity analysis shows that 29 of 36 nearby detector settings preserve the qualitative result.

Unsupported claims include clinical readiness, diagnostic safety, real hospital federated deployment performance, universal detector calibration, institutional ranking or institutional reliability judgment, and the claim that any real hospital, pathologist, or institution is inherently more or less trustworthy than another.

17 Reproducibility artifacts

The result tables, run diagnostics, diagnostic summaries, ablation outputs, calibration-sensitivity outputs, and figure-generation scripts are released in the project repository. Long filesystem paths are shortened in the paper body; exact paths are available from the repository's reproducibility documentation and source tree.

- Detector summary CSV and per-run diagnostics.
- Diagnostic-frequency summary by stress regime.
- Leave-one-diagnostic-family-out ablation table.
- Calibration-sensitivity sweep table.
- Figure-generation scripts for the schematic and paper figures.

Repository: https://github.com/matthewvaishnav/computational-pathology-research.

18 Conclusion

FedAvg is useful, but it silently equates sample count with aggregation authority. In pathology, sample volume and task-specific site-signal alignment can diverge. I simulate that divergence by perturbing the dominant client's training signal while keeping validation labels clean. Under conservative ordinal threshold shift, a fixed label-noise-calibrated detector switch keeps clean-regime switching low and improves global QWK, macro-F1, and worst-site QWK at 35% and 45% shift. Diagnostic ablation and calibration-sensitivity analysis suggest that the detector is not merely a one-feature or one-threshold artifact. The result supports sample-volume / site-signal alignment auditing as a research direction for federated computational pathology.

References

[1] Wouter Bulten et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28:154–163, 2022.

[2] Alexandre Filiot et al. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023.

[3] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.