

Censoring-Aware Rank Aggregation of Federated Top-k Explanation Prefixes: A Seed-Blocked Empirical Characterization against Weighted Borda Baselines

Aritrik Ghosh

aritrikghosh012@gmail.com

swami vivekananda university <https://orcid.org/0009-0003-7016-3315>

Research Article

Keywords: federated learning, explainable AI, rank aggregation, partial rankings, Plackett–Luce

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11339510/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Censoring-Aware Rank Aggregation of Federated Top- k Explanation Prefixes: A Seed-Blocked Empirical Characterization against Weighted Borda Baselines

Aritrik Ghosh

ARITRIKGHOSH012@GMAIL.COM

*Department of Computer Science and Engineering
Swami Vivekananda University
Kolkata, West Bengal, India*

Abstract

Clients in federated learning may disclose only ordered top- k feature-importance prefixes, leaving the tail unordered. We compare an established prefix-censored Plackett–Luce likelihood fitted by minorization–maximization with sample-size-weighted Borda on the UCI Heart Disease benchmark with five simulated clients. The primary analysis endpoint is Kendall agreement with a centralized reference at $k = 10$, averaging paired differences across four model families within each of ten seed blocks. The mean difference (censored PL minus Borda) is -0.0052 (tie-aware exact signed-rank $p = 0.453$; 95% t interval $[-0.0187, 0.0084]$), providing no evidence of improved agreement. NDCG@10 favors Borda in all ten seed blocks and remains significant after Holm correction across 11 secondary tests ($p_{\text{adj}} = 0.0215$); it remains exploratory because it was not prospectively registered. Prefix-resolved pair coverage is 45.9%, 61.2%, and 87.2% at $k = 3, 5, 10$. Synthetic findings are generator-specific. Higher-budget explanations, stronger baselines, and end-to-end reproduction remain outstanding.

Keywords: federated learning, explainable AI, rank aggregation, partial rankings, Plackett–Luce

1 Introduction

Federated learning (FL) trains a shared model from decentralized data by aggregating locally computed updates rather than centralizing raw records (McMahan et al., 2017). In healthcare, this architecture is attractive because patient-level data are difficult to pool across institutions. Explainability introduces an additional distributed object: clients may produce feature-importance signals that a server wishes to summarize without receiving complete attribution vectors. This creates a tension between information disclosure and aggregation quality. Prior work has shown that model explanations can expose information about training data, and recent surveys emphasize that the attack surface depends on the explanation signal and the adversary’s access path (Shokri et al., 2021; Zhao et al., 2021; Saifullah et al., 2024).

A client that releases only its top- k features does not report a complete permutation. The disclosed prefix resolves the first k positions, but the remaining features are only known to lie below the disclosed prefix and are not internally ordered. Generic rank aggregation can still be applied, but treating the censored tail as if it were a tied positional block discards the sequential-choice interpretation of the observation. Fed-CENSOR instead treats the released

prefix as a partial ranking under a Plackett-Luce model (Luce, 1959; Plackett, 1975; Turner et al., 2020).

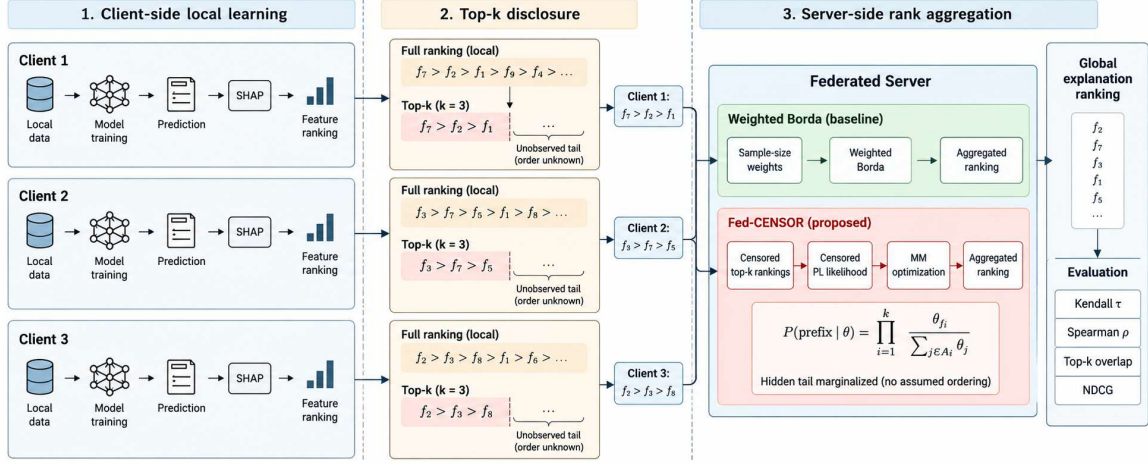


Figure 1: Fed-CENSOR workflow. Clients train locally, construct explanation rankings, disclose only a top- k prefix, and the server aggregates the partial rankings with Weighted Borda or the censored Plackett–Luce estimator.

The central scientific question is whether explicitly modeling the unresolved tail improves ranking recovery enough to justify the added assumptions and fitting cost over a positional baseline. This is an empirical comparison of established aggregation approaches, not a claim that the Plackett–Luce likelihood or its MM estimator is new. This study therefore treats Weighted Borda as a serious baseline, uses a centralized reference ranking as an evaluation anchor, and analyzes when the methods diverge rather than optimizing only for an average score. The study has three bounded objectives:

1. Observation model. We represent disclosed client explanations as ordered prefixes with an unresolved tail under an established Plackett–Luce likelihood.

2. Dependency-aware comparison. We compare censored PL-MM with sample-size-weighted Borda using seed-level paired inference, a centralized reference, disclosure sensitivity, and tie-aware evaluations.

3. Conditional-use characterization. We describe how observed differences depend on disclosure level, synthetic data-generating assumptions, and score/tie conventions.

We do not claim novelty for Plackett–Luce, MM estimation, Borda aggregation, top- k rankings, or partial-ranking likelihoods.

1.1 Research questions

RQ1. Does censoring-aware Plackett–Luce aggregation differ from sample-size-weighted Borda in agreement with a centralized reference under partial disclosure?

RQ2. How does the disclosure budget affect reference agreement and the amount of pairwise order information that remains unresolved?

RQ3. How do client heterogeneity, client count, and the model-family/federation mechanism relate to the observed aggregation results? These comparisons are descriptive, not causal, because the implementation uses different federation mechanisms for linear/MLP and tree models.

RQ4. How sensitive are conclusions to client disagreement, Borda tail-score conventions, alternative score representations, and numerical convergence?

2 Related Work

2.1 Federated learning and healthcare

Federated learning (FL) aggregates locally computed model information without centralizing raw training records (McMahan et al., 2017). Surveys and healthcare studies have documented both the promise and the open methodological challenges of cross-silo FL (Kairouz et al., 2021; Rieke et al., 2020). Healthcare-oriented explanation aggregation has also been studied, including federated XAI case studies and explainable federated systems with security-oriented goals (Ducange et al., 2024; Zhao et al., 2024). These works motivate the setting of decentralized explanation sharing, but they do not by themselves establish a new rank-aggregation estimator for the partial-ranking observation considered here.

2.2 SHAP and explanation disclosure

SHAP provides model-attribution values that can be summarized into feature-importance rankings (Lundberg and Lee, 2017). Explanation artifacts can also be informative to an adversary: prior work studies privacy risks and model-inversion attacks that exploit explanation signals (Shokri et al., 2021; Zhao et al., 2021), while recent work emphasizes that risk depends on how the explanation is acquired and what the adversary already knows (Saifullah et al., 2024; Oksuz et al., 2026). These results motivate caution when interpreting explanation disclosure as privacy-preserving. This paper does not evaluate privacy protection and makes no differential-privacy claim.

2.3 Rank aggregation, partial rankings, and Plackett–Luce

Borda is a classical positional scoring rule. Partial-ranking aggregation has a broader literature spanning top- k lists, rank-breaking, probabilistic ranking models, and model-robust procedures (Fagin et al., 2003; Chen et al., 2022; Soufiani et al., 2014; Khetan and Oh, 2016, 2018; Pan et al., 2022). The Plackett–Luce (PL) model itself is classical (Luce, 1959; Plackett, 1975), and established software and theory cover partial-ranking likelihoods and related estimation issues (Turner et al., 2020). Ma et al. consider partitioned preference, in which items are divided into ordered partitions with unknown within-partition order, and identify top- k ranking as a special case under PL (Ma et al., 2021). The observation structure used here therefore substantially overlaps with established partial-ranking PL methodology. Broader models can also combine rankings with other information; for example, Pearce and Erosheva study a unified ranking-and-score model rather than the top- k observation used here (Pearce and Erosheva, 2022).

Note. “Yes” indicates direct study of the property; “Related” indicates adjacent methodology rather than the same observation. Turner et al. and Ma et al. directly cover the top- k PL prefix observation,

Table 1: Positioning of Fed-CENSOR relative to closely related ranking work.

Work	Partial	Top- k	Censor.	PL	Fed-XAI	MM	Same obs.
Plackett (1975)	–	–	–	Yes	–	–	–
Hunter (2004)	–	–	–	–	–	Yes	–
Soufiani et al. (2014)	–	–	–	Yes	–	–	–
Khetan–Oh (2016)	Related	–	–	Related	–	–	–
Khetan–Oh (2018)	Related	–	–	Yes	–	Yes	–
Turner et al. (2020)	Yes	Yes	–	Yes	–	–	Yes
Ma et al. (2021)	Yes	Yes	–	Yes	–	–	Yes
Pan et al. (2022)	Related	–	Related	Yes	–	Related	–
Chen et al. (2022)	Yes	Yes	–	–	–	–	–
Migali et al. (2026)	Yes	Yes	Yes	Yes	–	–	Related
Fed-CENSOR	Yes	Yes	Yes	Yes	Yes	Yes	Application-specific

while Migali et al. present a broader incomplete-ranking/positional-censoring framework. None of these works evaluates federated explanation aggregation in the present setting.

Recent work broadens the problem further. Migali et al. study learning from incomplete rankings under a general feedback framework and introduce positional censoring, which encompasses standard forms of partial feedback while allowing a wider class of ranking models (Migali et al., 2026). Pan et al. study robustness to model misspecification and instantiate CoarsenRank for several probabilistic rank models, including Plackett–Luce (Pan et al., 2022). These papers narrow the novelty space for any claim based solely on “censoring” or “partial PL”.

2.4 Position of this study

The contribution is an empirical characterization of an established top- k partial-ranking likelihood in federated explanation sharing. We test whether it changes agreement relative to weighted Borda, distinguish observed prefix relations from relations created by a deterministic tail tie-break, and report when conclusions depend on the score representation or disclosure budget. The evidence is bounded to simulated clients and retained ranking payloads; it does not establish a new ranking estimator or universal superiority.

No priority claim is made for the underlying Plackett–Luce likelihood, MM estimation, Borda aggregation, or top- k partial rankings.

3 Problem Definition

Let a client rank F features by decreasing explanation importance. Its latent complete permutation is $\pi=(i_1, \dots, i_F)$, but the server receives only the ordered prefix $y=(i_1, \dots, i_k)$. The observation resolves the first k positions, establishes that every undisclosed feature is below the disclosed prefix, and leaves the internal order of the remaining $F-k$ features unknown. For non-negative Plackett–Luce strengths $\theta = (\theta_1, \dots, \theta_F)$, normalized so that the strengths sum to 1, the censored prefix likelihood is

$$\Pr(i_1, \dots, i_k \mid \theta) = \prod_{s=1}^k \frac{\theta_{i_s}}{\sum_{j \in R_s} \theta_j}. \quad (1)$$

where R_s is the complete set of features that remain available at stage s . Thus the undisclosed tail remains in the denominators whenever it is still eligible, but its internal

order is not invented. This is the statistical interpretation used by Fed-CENSOR; it is not a privacy definition. The likelihood is the standard top-k Plackett–Luce prefix likelihood; the paper’s methodological distinction is its use as an application-specific observation model for federated explanation disclosure rather than a new PL likelihood.

4 Aggregation Procedures

4.1 Client-side disclosure

The client first converts mean absolute SHAP magnitude into a descending feature ranking. The disclosed object is the ordered prefix $y = (i_1, \dots, i_k)$. All features outside the prefix are known only to occur below the disclosed items; their internal order is not observed. For implementation and storage, the undisclosed features are represented by a common cutoff rank. Whenever a downstream metric requires a strict permutation, a stable feature-index sort is used only to break those artificial ties. That deterministic order is an implementation convention and is not interpreted as evidence about the latent tail.

4.2 Weighted Borda baseline

For client c with rank vector $r(c)$, Borda uses the positional score

$$s_j^{(c)} = (F - 1) - r_j^{(c)}. \quad (2)$$

Client scores are multiplied by normalized sample-size weights and summed. Features are then ordered by descending aggregate score. This is a deterministic, non-parametric baseline operating on exactly the same disclosed payloads as Fed-CENSOR. The wider top-k Borda literature includes theoretical analysis of selection from partial rankings (Chen et al., 2022). For the primary retained implementation, every undisclosed feature receives zero-based rank k (the next unseen rank) before Borda scoring; this creates a common-score block among features omitted by every client. This is one truncated-Borda convention, not an observed order of the hidden tail. To assess how much the convention matters, we also recompute Borda using the last disclosed rank, the mean possible tail rank, and the bottom rank on the same retained prefixes (Appendix Table A15). These are sensitivity analyses, not model retraining, and none is selected by optimizing against the centralized reference.

4.3 Censored Plackett-Luce MM estimator

For client c , define the weighted win count

$$W_j = \sum_c w_c \sum_{s=1}^k \mathbf{1}\{i_s^{(c)} = j\}. \quad (3)$$

At iteration t , define the exposure term

$$D_j(\theta^{(t)}) = \sum_c w_c \sum_{s=1}^k \mathbf{1}\{j \in R_{cs}\} \frac{1}{\sum_{\ell \in R_{cs}} \theta_\ell^{(t)}}. \quad (4)$$

The implemented MM fixed-point update follows the standard MM construction for Plackett–Luce-type worth models (Hunter, 2004):

$$\tilde{\theta}_j^{(t+1)} = \frac{W_j}{\max\{D_j(\theta^{(t)}), 10^{-12}\}}, \quad \theta_j^{(t+1)} = \frac{\tilde{\theta}_j^{(t+1)}}{\sum_{\ell} \tilde{\theta}_{\ell}^{(t+1)}}. \quad (5)$$

The implementation permits $\theta_j=0$ when $W_j=0$. Convergence is declared when the maximum absolute coordinate change is below the configured tolerance. The implementation does not independently verify graph connectivity or use a likelihood-improvement stopping criterion.

4.4 What the model assumes

Fed-CENSOR assumes that the disclosed prefix can be represented by a common non-negative worth vector after client weighting. It does not model client-specific worth parameters, ranker reliability, or systematic reporting bias. If clients strongly disagree, the fitted worth vector is a compromise under the Plackett-Luce likelihood. If a feature is never disclosed, its empirical win count is zero and the implemented update places its fitted strength on the boundary. Consequently, the relative ordering among never-disclosed features is not statistically identified by observed wins alone and should not be interpreted as learned evidence.

4.5 Finite-estimate and connectivity considerations

The likelihood is defined for non-negative worth parameters, but a finite interior maximum is not automatic for arbitrary partial-ranking data. A standard diagnostic in Plackett–Luce estimation is the connectivity of the induced win/loss comparison structure: when the observed comparisons leave a subset of features insufficiently connected to the remainder, one or more worth parameters can be driven toward the boundary rather than identifying a unique interior solution. Modern Plackett–Luce implementations use explicit network diagnostics and, in some settings, pseudo-comparisons to obtain finite estimates (Turner et al., 2020). Fed-CENSOR does not add pseudo-comparisons. We therefore treat comparison connectivity as an identifiability condition that should be checked for a deployed fit. The present release records the numerical stopping rule and boundary behavior, but it does not contain an independent connectivity audit for every retained fit, so we do not claim a finite interior MLE for every dataset.

4.6 Relation to Borda

Borda and Fed-CENSOR can produce nearly identical rankings when the disclosed prefixes are already highly concordant. A simple special case is complete agreement of all client prefixes: both methods preserve the common order of the disclosed items, while the unobserved tail is unresolved by the data. More generally, Borda is a positional summary, whereas Fed-CENSOR uses the sequential-choice denominators contributed by all remaining features. Divergence is therefore possible when the same features are repeatedly selected at different positions across clients and the probabilistic likelihood assigns different relative worth than a linear positional score.

4.7 Complexity

Let C be the number of clients, F the feature count, k the disclosed prefix length, and T the number of MM iterations. The direct implementation scans the remaining features at each stage, giving $O(TCkF)$ aggregation work. Because each denominator can be represented as the remaining total worth after removing previously selected items, cumulative-prefix sums reduce the per-iteration cost to $O(C(k + F))$ up to sorting and normalization constants. A cumulative-sum implementation was benchmarked post hoc; it reduced median Fed-CENSOR routine time by approximately 2.40x at $k = 3$ and 3.59x at $k = 10$ while leaving parameter differences at machine precision. Memory is $O(F + Ck)$ for the fitted parameters and disclosed sequences, excluding model state and retained round histories.

5 Experimental Protocol

Table 2. Experimental program and evidence hierarchy.

Study	Core design	Reference	Evidence role
Historical factorial	5 seeds; 3/5/8 clients; $\alpha = 0.1/0.5/1.0$; $k = 3, 10$; four models	Legacy agreement	Historical reanalysis
Confirmatory	10 seeds; 5 clients; $\alpha = 0.5$; $k = 3, 10$; four models	Centralized reference	Primary inferential study
Frozen extension	Seeds 101–120; $k = 10$; four models	Centralized reference	Seeds 111–120 not row-level auditable
Prefix-resolved pairs	10 confirmatory seeds; top- k pair mask	Centralized reference	Derived sensitivity
Synthetic ground truth	3 regimes; $F = 28$, $C = 10$; $k = 3, 5, 10, 28$; 100 replicates/condition	Generating worth	Executed validation
Full-ranking ablation	Seeds 121–130; masked/full rankings; $k = 3, 10, 28$	Centralized reference	Executed ablation
Heterogeneity audit	Seeds 101–110; realized client-class proportions	Partition output	Descriptive design audit
External validation	Seed 301; WDBC and Diabetes; $k = 1, 3, 5, 10$	Centralized reference	Secondary descriptive

5.1 Dataset and preprocessing

The benchmark is the UCI Heart Disease dataset assembled from four source cohorts: Cleveland (303 records), Hungary (294), Switzerland (123), and VA (200), for 920 raw records in the supplied source snapshot (Janosi et al., 1989). The original clinical study associated with this benchmark is reported by Detrano et al. (Detrano et al., 1989). The pipeline uses five numeric and eight categorical variables, median/mode imputation, numeric standardization, and one-hot encoding, producing 28 transformed features. The binary target is $\text{num} > 0$. The confirmatory study uses a fixed split of 644 training, 92 validation, and 184 test records. Imputation, scaling, and one-hot encoding parameters are fit on the training partition only and then applied unchanged to validation and test data. The training data are partitioned into simulated clients with class-wise Dirichlet allocation followed by a repair

step guaranteeing at least two observations of each binary class per client. The realized positive-class proportions across the ten confirmatory seeds ranged from approximately 0.008 to 0.988, showing substantial residual label skew after repair. This is therefore a constrained/repaired Dirichlet construction rather than a pure unconstrained Dirichlet draw.

5.2 Federated models

The study uses logistic regression, MLP, Random Forest, and XGBoost. Logistic regression uses an SGD logistic-loss model with $\alpha = 10^{-4}$ and learning rate 0.01; the MLP has one hidden layer of 32 units and learning-rate initialization 0.01 with warm starts; Random Forest uses 100 trees; XGBoost uses 100 estimators with learning rate 0.01. These model settings are fixed benchmark configurations rather than results from a model-selection search; XGBoost in particular is intentionally conservative and should not be interpreted as a tuned predictive benchmark. Logistic regression and MLP use parameter FedAvg. Random Forest and XGBoost do not use parameter-space FedAvg; local tree models are retrained and server predictions are combined by a sample-size-weighted probability ensemble. This distinction is preserved throughout interpretation of model-family results.

5.3 Explanation construction

The historical source implementation uses SHAP through the client explainer path, but exact historical SHAP background/evaluation counts recorded in legacy manifests are not authoritative runner parameters. The confirmatory implementation uses five background samples, three client explanation samples, and 32 KernelSHAP samples for MLP; it uses LinearExplainer for logistic regression and TreeExplainer for Random Forest/XGBoost. For multi-output SHAP values, the final positive-class output is selected, absolute attribution magnitudes are taken, and values are averaged over the evaluation samples before ranking. These confirmatory settings are not retroactively assigned to the historical 720 rows.

5.4 Historical factorial study

The historical result table contains 720 completed logical run specifications: five seeds x three client counts x three Dirichlet concentrations x two disclosure levels x four model families x two aggregation methods. There are 180 unique seed x client-count x alpha x model cells. The runner creates separate result specifications for top-k and aggregation method, so the release does not reinterpret the 180 cells as 720 independent physical training trajectories. Historical training uses 15 rounds, two local epochs, 100 tree estimators, and sample-size weighting.

5.5 Confirmatory centralized-reference study

The confirmatory study uses seeds 101–110, five simulated clients, $\alpha = 0.5$, 15 rounds, two local epochs, sample-size weighting, SHAP-based rankings, and $k \in \{3, 10\}$. Weighted Borda and Fed-CENSOR operate on the same disclosed client rank payload within each seed/model trajectory. Separately trained centralized models provide a model-derived reference ranking; this reference does not participate in federation or aggregation and is not treated as absolute feature-importance ground truth. There are 40 centralized predictive rows and 160 method-

level reference-comparison rows. The primary endpoint is centralized-reference Kendall’s tau at $k = 10$. The method difference is first averaged across the four model families within each seed, yielding ten independent seed blocks. Secondary endpoints are Spearman’s ρ , top-k overlap, Jaccard, normalized footrule, and NDCG@10. Secondary tests use Holm correction within the declared secondary family. Model-family summaries are descriptive because the four models use two different federation mechanisms.

5.6 Additional retained-data analyses

Five supplementary analyses use retained outputs and do not require fresh model training:

- Nested disclosure: retained top-10 client prefixes are truncated to $k=1$ and 5, in addition to 3 and 10. These are deterministic nested reconstructions from the retained top-10 payloads, not independent reruns.
- Additional baseline: a weighted top-k occurrence-count rule is implemented as a simple non-parametric baseline. It counts weighted appearances in the disclosed prefix and deterministically breaks ties by average disclosed position and feature index.
- Client disagreement: disagreement is defined as one minus pairwise top-k overlap among client disclosed sets. Association with the Fed-CENSOR minus Borda reference Kendall difference is analyzed after averaging over model families within seed; this is exploratory with only ten seed-level observations.
- MM numerical analysis: the same retained inputs are refit with maximum iterations 100, 500, and 1000 under tighter tolerance. Rank-vector stability is checked for the three cases that fail the 100-iteration criterion.

5.7 Evaluation metrics

All ranking metrics are computed against the centralized reference ranking for the real benchmark. Let $\hat{\pi}$ and π^{ref} denote strict permutations of the same F features and let r_j and r_j^{ref} denote their zero-based rank positions (smaller is better).

Kendall’s τ . Kendall’s τ is computed as the correlation between the two rank vectors, equivalently comparing concordant and discordant feature pairs. Positive values indicate greater agreement.

Spearman’s ρ . Spearman’s ρ is the Pearson correlation between the two rank vectors. Positive values indicate greater agreement.

Top-k overlap and Jaccard. With A_k and B_k denoting the sets of the first k items under the two rankings,

$$\text{Overlap}@k = \frac{|A_k \cap B_k|}{k}, \quad J@k = \frac{|A_k \cap B_k|}{|A_k \cup B_k|}. \quad (6)$$

Higher values indicate greater agreement.

Normalized footrule. We use the retained implementation

$$NF(\hat{\pi}, \pi^{\text{ref}}) = \frac{\sum_{j=1}^F |r_j - r_j^{\text{ref}}|}{F^2/2}. \quad (7)$$

where lower values indicate greater agreement.

NDCG@10. Let the candidate ordering select items $\hat{\pi}1, \dots, \hat{\pi}10$. The retained implementation maps a reference rank $\text{rref } j \in \{0, \dots, F-1\}$ to relevance $\text{gj} = F - \text{rref } j$ and uses exponential gain:

$$\text{DCG@10} = \sum_{s=1}^{10} \frac{2^{g_{\hat{\pi}_s}} - 1}{\log_2(s+1)}, \quad \text{NDCG@10} = \frac{\text{DCG@10}}{\text{IDCG@10}}. \quad (8)$$

where IDCG is obtained by ordering the same relevance values by decreasing reference relevance. Higher values indicate greater agreement. Ties in the candidate strict permutation are resolved by the stable feature-index convention described above. This secondary metric is fully specified by the retained implementation: relevance is $\text{gj} = F - \text{rref } j$, gains are exponential, the cutoff is 10, and IDCG uses the same reference-derived relevance values in descending order. It remains a metric against a model-derived reference, not an independent notion of feature-importance truth.

Prefix-resolved pair coverage and Kendall. A disclosed prefix identifies the ordering of its items and establishes that every omitted feature lies below every disclosed item. Therefore, for a given client, a pair is ordered whenever at least one member is disclosed: the pair is ordered within the prefix if both members appear, and prefix-versus-tail if only one appears. Across clients, we define the prefix-resolved mask to include every pair for which at least one feature appears in at least one retained client prefix. A pair whose two features are never disclosed by any client is excluded. For $F = 28$, coverage is $1 - \binom{U}{2} / \binom{28}{2}$, where U is the number of features never disclosed across clients. This mask means a relation is observed for at least one client; it does not imply unanimous client preference. Prefix-resolved Kendall agreement compares the aggregated ordering with the centralized reference only on these pairs. We separately report a narrower co-disclosed-pair sensitivity that counts only pairs appearing together inside a disclosed prefix; that statistic does not represent all information conveyed by censoring.

For inferential comparisons, the reported mean difference is always Fed-CENSOR minus Borda, and the independent unit is the seed block after averaging the paired difference across the four model families.

6 Historical Agreement Analysis

The historical Fed-CENSOR minus Weighted Borda differences are +0.000886 for Spearman rho, -0.000599 for Kendall tau, -0.00603 for top-k overlap, and +0.000832 for normalized footrule distance. Since larger footrule indicates greater rank displacement, its positive difference also does not favor Fed-CENSOR.

The independent unit is the five-seed block, not the repeated configuration rows. Lower footrule is better. The seed-blocked exact two-sided Wilcoxon test gives $p=0.0625$ for each of the four endpoints. Holm-adjusted values are 0.25. With only five independent seed blocks, the test has coarse resolution and 0.0625 is the smallest possible nonzero two-sided exact p-value for five nonzero differences. These results are not evidence of superiority. For normalized footrule, the seed-level mean is +0.000832 with a 95% t interval [0.000544, 0.001120]; a parametric one-sample t sensitivity gives $p \approx 0.0013$ ($df=4$). The direction is unfavorable to Fed-CENSOR because larger footrule means greater rank displacement. The exact Wilcoxon result remains $p = 0.0625$ because only five seed blocks are available

Table 4. Historical seed-blocked Fed-CENSOR minus Weighted Borda differences. Five independent seed blocks; confidence intervals are t intervals.

Endpoint	Mean difference	Exact p	95% CI	Cohen’s d_z	Holm p
Spearman ρ	+0.000886	0.0625	[-0.000022, 0.001793]	+1.212	0.25
Kendall τ	-0.000599	0.0625	[-0.001277, 0.000079]	-1.096	0.25
Top- k overlap	-0.006033	0.0625	[-0.013347, 0.001281]	-1.024	0.25
Normalized footrule	+0.000832	0.0625	[0.000544, 0.001120]	+3.590	0.25

and the exact test has coarse resolution. The historical data also show a large disclosure effect. Mean Spearman agreement rises from approximately 0.4835/0.4845 for Weighted Borda/Fed-CENSOR at $k=3$ to 0.7232/0.7241 at $k=10$. The method difference is therefore orders of magnitude smaller than the disclosure-level change. Mean normalized footrule decreases from about 0.778 to 0.468 as k increases.

7 Theoretical and Numerical Analysis

7.1 Censoring and boundary identifiability

The censored likelihood uses the full remaining feature set in every denominator. This is the key statistical distinction from an approach that pretends the undisclosed tail has a known internal order. At the same time, the retained data show that many features are never disclosed. Across the 40 seed/model blocks, the mean number of zero-win features is about 20.6 at $k=3$ and 10.0 at $k=10$. Thus a large part of the 28-feature tail can remain empirically unresolved, especially at low disclosure. This provides a concrete explanation for why a more expressive probabilistic model need not substantially alter the global ranking: the information needed to distinguish many tail features is simply absent.

7.2 Disclosure curve

Using the retained top-10 payloads, the nested sample-weighted reference analysis gives the following mean Kendall tau values:

Derived disclosure curve from retained top-10 payloads.			
k	Weighted Borda Kendall	Fed-CENSOR Kendall	Difference
1	0.3647	0.3647	0.0000
3	0.4041	0.4040	-0.0001
5	0.4450	0.4474	+0.0024
10	0.5218	0.5164	-0.0054

The same analysis shows increasing mean centralized-reference agreement as more of the client prefix is disclosed, while the sign of the method difference changes with k . These are nested truncations of the same retained payloads, so the disclosure curve is derived rather

than a set of independent experiments. The $k = 1$ and $k = 5$ results are derived from retained top-10 client prefixes and should not be interpreted as independent training experiments.

7.3 Does client disagreement predict when Fed-CENSOR helps?

The retained-data analysis defines client disagreement as one minus pairwise top- k overlap. At $k = 3$, the seed-level Spearman association between disagreement and the Fed-CENSOR minus Borda Kendall difference is 0.317 ($p = 0.372$); at $k = 10$ it is -0.328 ($p = 0.354$). Kendall associations are similarly non-significant. With only ten seed-level observations, these analyses have limited power and are exploratory. The available data therefore do not identify a reliable monotone condition under which censored-likelihood aggregation systematically outperforms Weighted Borda.

7.4 MM convergence and rank stability

At the default 100-iteration budget, 39/40 primary censored-MM fits meet the stopping rule at $k = 3$ and 38/40 at $k = 10$. The two $k = 10$ fits that do not meet the criterion are retained in the original primary rank comparison, so that result should not be interpreted as evidence of parameter convergence for every fit. In a longer-run sensitivity (1,000 iterations, tolerance 10^{-8}), the rank vectors from the originally non-converged fits remain unchanged; this supports ranking stability for those cases but does not retroactively label the original fits as converged. A separate convergence-gated baseline audit shows RankBreak-PL-MM converged in 7/40 fits at $k = 3$, 12/40 at $k = 5$, and 21/40 at $k = 10$. Regularized BT-LBFGS converged in 40/40 fits at each of $k = 3, 5$, and 10. At $k = 1$, pairwise baselines are not identifiable. Failed/non-converged scores are not treated as valid estimates. These audit layers use separate implementations and are reported separately.

regularized Bradley–Terry LBFGS optimizer audit.

Diagnostic	Observed result
Optimizer termination	120/120 runs reported optimizer success
Runs meeting requested gtol ($1e-9$)	53/120
Runs meeting review threshold ($1e-7$)	120/120
Maximum gradient infinity norm	4.176315879e-8
Representative finite-difference gradient error	2.047352912e-10
Interpretation	Optimizer status is reported separately from strict gtol attainment; this is not the rank-breaking PL-MM convergence audit.

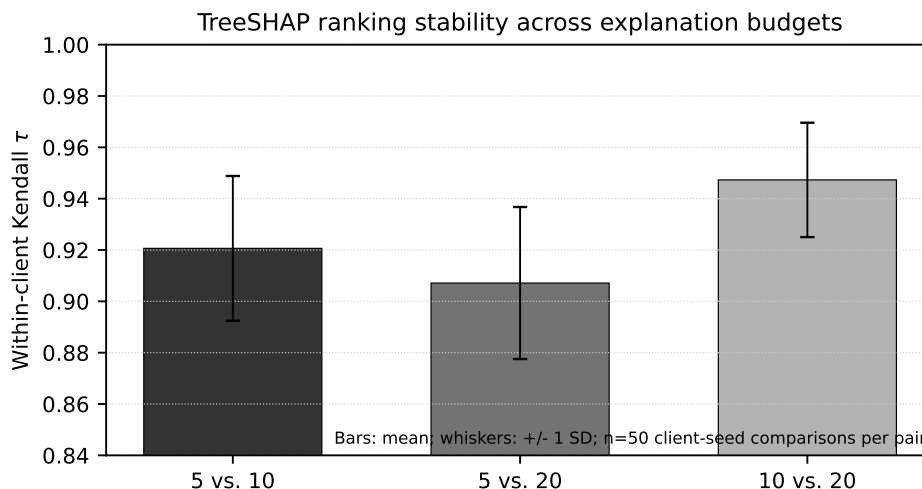
7.5 Executed explanation-budget sensitivity

The exploratory explanation-budget sensitivity experiment uses the Breast Cancer Wisconsin Diagnostic dataset (30 features), RandomForest TreeSHAP, and SHAP 0.52.0. Each of ten seeds (101-110) and five simulated clients is scored at explanation budgets of 5, 10, and 20 examples, yielding 50 paired seed-client comparisons for each budget pair. Budgets refer to

the number of examples used to estimate local attribution scores, not the number of features disclosed.

Within-client WDBC TreeSHAP stability across explanation budgets.

Budget pair	Pairs	Mean Kendall tau (SD)	Mean top-10 overlap
5 vs. 10	50	0.9206 (0.0282)	0.972
5 vs. 20	50	0.9071 (0.0296)	0.972
10 vs. 20	50	0.9473 (0.0223)	0.980



Supplementary figure. Within-client rank stability for WDBC RandomForest–TreeSHAP across explanation-example budgets 5, 10, and 20.

These results indicate high but imperfect within-client ranking stability as the explanation sample budget changes. They do not demonstrate invariance, do not isolate every source of SHAP estimation error, and do not replace the retained Heart Disease primary analysis. This separate WDBC TreeSHAP study does not validate budget sensitivity for the LinearExplainer or KernelSHAP paths used by the Heart Disease model families.

8 Confirmatory Centralized-Reference Evaluation

The centralized reference is an independently trained, model-derived evaluation anchor, not ground-truth feature importance. Both aggregators receive identical disclosed client prefixes; the method difference is assessed after averaging the paired differences across four model families within each of ten independent seed blocks.

For the primary endpoint used in the confirmatory comparison, the ten seed-block differences in centralized-reference Kendall’s τ at $k = 10$ have mean -0.00516 , tie-aware exact signed-rank $p = 0.453$, Cohen’s $d_z = -0.272$, and a 95% t interval $[-0.01873, 0.00841]$. Five seed blocks favor each method. The primary comparison therefore does not establish an improvement. The magnitude is roughly one net discordant-pair equivalent on a 378-pair scale, although the reported quantity is an average of seed-level rank correlations rather than a literal pooled pair count. The tie-aware exact test uses average midranks for tied

absolute differences and enumerates all 2^{10} sign assignments. An auxiliary convergence-gated reconstruction using the retained ranking payloads gives a consistent conclusion: after excluding the one fit per k that fails the extended 5,000-iteration/ 10^{-8} criterion and pairing Borda on the same rows, the reconstructed strict-Kendall difference is $+0.00013$ at $k = 3$ (exact $p = 0.984$) and -0.00445 at $k = 10$ (exact $p = 0.604$; 95% t interval $[-0.01863, 0.00973]$). This is an auxiliary audit rather than a replacement for the primary endpoint because it uses reconstructed client weights and the independent payload audit retains one unresolved Borda-row discrepancy (seed 101, Random Forest, $k = 10$); detailed results are in Appendix Table A16.

At $k = 10$, the secondary differences are mixed: Spearman $\rho = -0.005624$, Jaccard $= -0.015418$, overlap $= -0.010000$, footrule $= +0.004209$, and NDCG@10 $= -0.021974$. The NDCG contrast favors Borda in all ten seed blocks (exact $p = 0.001953$; Holm-adjusted $p = 0.021484$ across the 11 secondary contrasts). Because relevance is exponentially weighted by centralized-reference rank, NDCG@10 emphasizes the top of the reference ranking. This result is therefore substantively relevant to top-feature recovery and should not be dismissed as merely metric-specific. However, it is a secondary, non-prospectively registered endpoint and is suggestive rather than decisive. It does not replace the primary endpoint.

At $k = 3$, the Kendall difference is zero at the retained precision. Direct method-to-method agreement is also high: mean direct Spearman/Kendall are approximately 0.998/0.989 at $k = 3$ and 0.982/0.933 at $k = 10$. Thus the two procedures often return similar global rank vectors from the same partial inputs.

Table 5. Centralized-reference inference. Paired differences are first averaged over the four model families within each seed. The primary endpoint is separate from the 11 secondary contrasts.

k	Metric	Mean difference	Exact p	Holm p
3	Kendall τ	0.000000	1.000000	1.000000
3	Spearman ρ	+0.000698	0.556641	1.000000
3	Top- k overlap [‡]	+0.008333	0.812500	1.000000
3	Jaccard [‡]	+0.007500	0.812500	1.000000
3	Footrule	-0.000255	0.843750	1.000000
3	NDCG@10	+0.000495 [‡]	1.000000 [‡]	1.000000 [‡]
10	Kendall τ (primary)	-0.005159	0.453125 [†]	—
10	Spearman ρ	-0.005624	0.322266	1.000000
10	Jaccard	-0.015418	0.203125	1.000000
10	Top- k overlap	-0.010000	0.148438	1.000000
10	Footrule	+0.004209	0.232422	1.000000
10	NDCG@10	-0.021974	0.001953	0.021484

Positive differences favor censored PL for correlation/overlap measures; lower footrule is better. Negative NDCG@10 difference favors Borda. Holm correction covers the six $k = 3$ metrics and the five non-primary $k = 10$ metrics (11 tests); the $k = 10$ primary endpoint is excluded. [†] Tie-aware exact signed-rank result using average midranks for tied absolute differences. [‡] Derived from retained per-client ranking payloads and reconstructed weights; these $k = 3$ secondary set/top-weighted metrics do not rerun training or SHAP.

9 Additional Validation of the Aggregation Question

These analyses address three distinct issues: the fraction of pairwise order actually identified by the disclosed prefixes, the behavior of estimators under specified synthetic generators, and

the behavior of alternative baselines on the same retained prefixes. They are supplementary and do not replace the confirmatory primary endpoint.

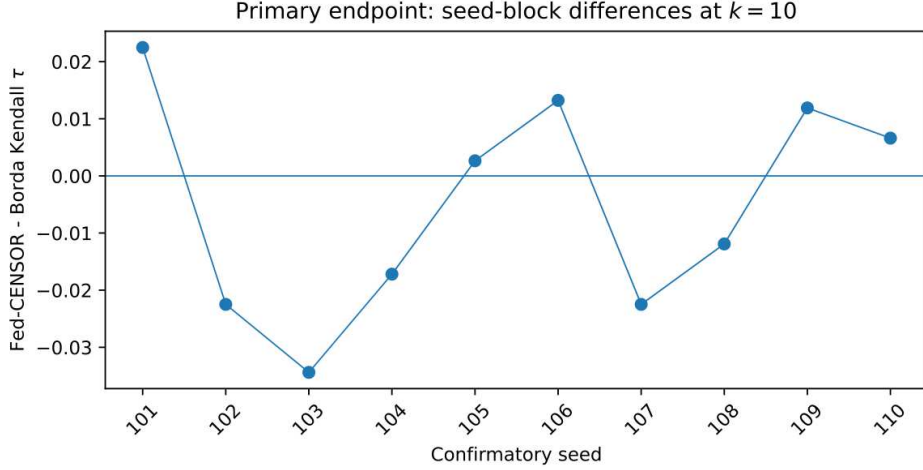


Figure 2: Seed-level Fed-CENSOR minus Borda differences for the primary centralized-reference Kendall endpoint at $k = 10$. Points show the ten independent seed blocks; positive values favor Fed-CENSOR.

9.1 Prefix-resolved pair coverage and agreement

For $F = 28$ features, there are $\binom{28}{2} = 378$ unordered pairs. Under the prefix-censoring observation model, the relative order of a pair is resolved for a client whenever at least one member is disclosed: both items are ordered within the prefix if both appear, and a disclosed item is known to outrank every omitted item. We therefore define total prefix-resolved coverage using the union of features disclosed by any client. The mean coverage is 45.9% at $k = 3$, 61.2% at $k = 5$, and 87.2% at $k = 10$ (Appendix Table A9 and Figure 4). The $k = 5$ result is a deterministic truncation of retained $k = 10$ payloads, not an independent training run. These coverage values do not imply that clients agree on every resolved pair; they describe which pair relations are observed for at least one client.

Using this prefix-resolved pair mask and restricting both methods to the same 39/40 model rows whose extended PL fits meet the convergence criterion, mean Kendall agreement with the centralized reference is 0.668, 0.639, and 0.573 for Weighted Borda at $k = 3, 5, 10$, respectively. The corresponding means for censored PL-MM are 0.668, 0.641, and 0.566 (Figure 3). These are descriptive matched-row summaries from reconstructed retained payloads, not seed-level inferential effects. Pair coverage itself uses all 40 payload blocks. A separate co-disclosed-pair statistic, which counts only pairs appearing together within a top- k prefix and therefore omits prefix-versus-tail relations, is retained as a narrower sensitivity in Appendix Table A13. The two masks answer different questions and should not be treated as interchangeable.

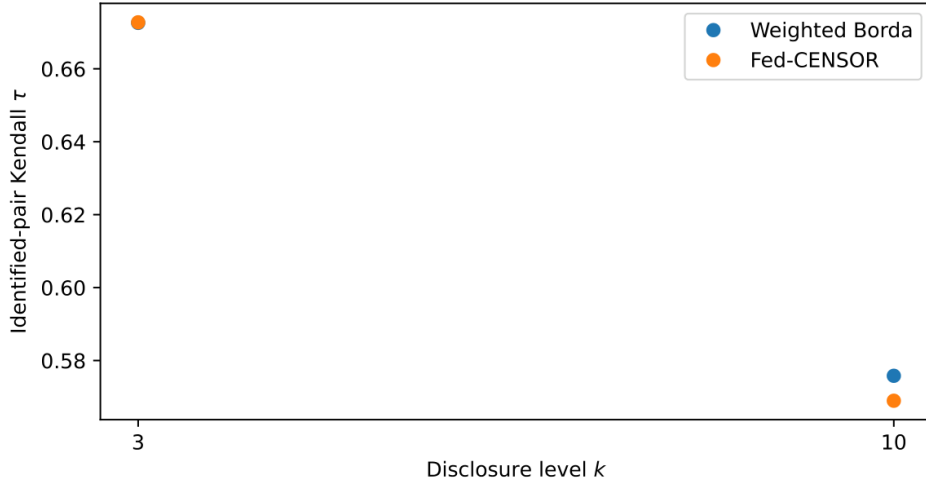


Figure 3: Prefix-resolved pair Kendall agreement against the centralized reference. A pair is included if at least one member appears in any client prefix. PL-MM means use valid fits (39/40 rows per k); points are descriptive, not paired-effect estimates.

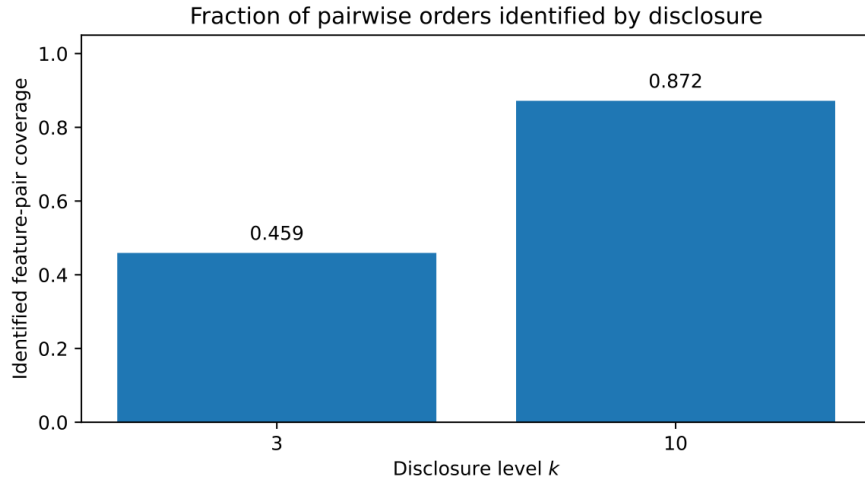


Figure 4: Fraction of all 378 unordered feature pairs resolved by retained client prefixes. A pair is included if at least one member is disclosed by at least one client; coverage is 45.9%, 61.2%, and 87.2% at $k = 3, 5, 10$.

9.2 Ground-truth synthetic Plackett–Luce validation

A separate simulation uses $F = 28$ features and $C = 10$ clients, with 100 independent replicates per regime and disclosure condition and $k \in \{3, 5, 10, 28\}$. In the common-PL condition, all clients sample rankings from a shared non-negative worth vector, so the generating worth order is known. The retained common-PL mean Fed-CENSOR-minus-Borda Kendall differences are -0.01222 , -0.00831 , $+0.01767$, and $+0.02720$ at $k = 3, 5, 10, 28$. The sign therefore varies with disclosure. These are results under correct model specification, not evidence of real-benchmark superiority.

Two retained stress regimes introduce client heterogeneity and rank noise. The artifact summary does not preserve the exact perturbation magnitudes needed to reconstruct a unique calibrated generator. We therefore report their mean differences as descriptive stress-test results only: heterogeneity, -0.01286 , -0.00550 , $+0.01138$, and -0.00185 ; rank noise, -0.01111 , -0.00063 , $+0.02143$, and $+0.01968$, at $k = 3, 5, 10, 28$, respectively. These results are not calibrated generative-model inference.

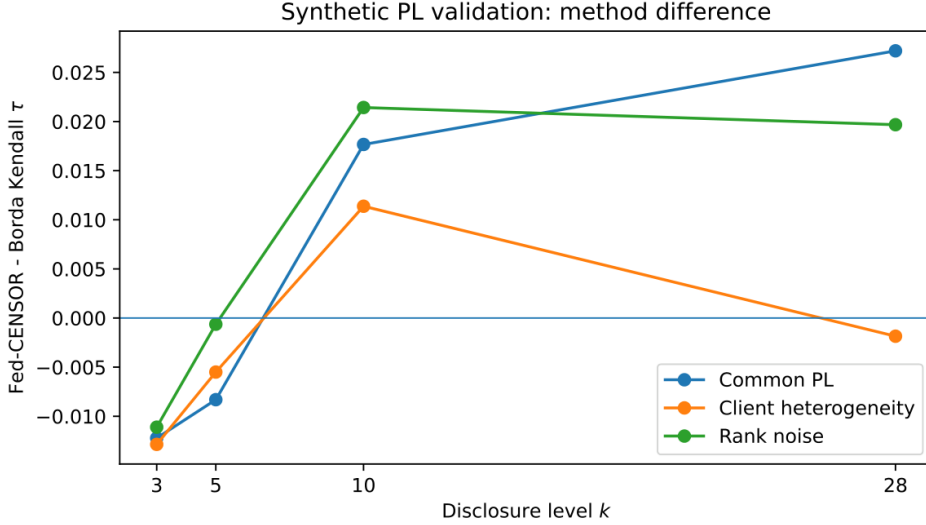


Figure 5: Ground-truth synthetic validation. Fed-CENSOR-minus-Borda Kendall differences vary with disclosure level and ranking-generation regime. Results are specific to the retained generators and do not establish real-data superiority.

9.3 Two-component mixture-PL misspecification stress test

A separate fully specified simulation creates two related worth vectors. For each replicate, 28 base worths are formed by normalizing $\exp(z_j)$, with $z_j \sim N(0, 0.72)$. A second worth vector multiplies each base worth by $\exp(\epsilon_j)$ and renormalizes, with the recorded perturbation distribution $\epsilon_j \sim N(0, 1)$. Each of ten clients independently samples either component with probability 0.5 and discloses only its top k features. The target is the descending Copeland score induced by the mixture’s pairwise PL win probabilities. This is a model-defined population order, not a universal consensus target.

Table 6: Synthetic mixture-PL stress test. Mean full-ranking Kendall against the population Copeland target over 100 paired simulation replicates per k .

k	Borda	Occurrence	Rank-break PL	Fed-CENSOR	Copeland
3	0.36534	0.37317	0.37153	0.37291	0.37090
5	0.46942	0.48598	0.48471	0.48593	0.48556
10	0.56847	0.59725	0.59582	0.59979	0.59672
28	0.71810	0.71810	0.71794	0.72894	0.71815

For this generator, the Fed-CENSOR-minus-Borda mean deltas are $+0.00757$ at $k = 3$ (95% paired bootstrap interval $[0.00381, 0.01148]$), $+0.01651$ at $k = 5$ $[0.01079, 0.02249]$, $+0.03132$ at $k = 10$ $[0.02275, 0.04000]$, and $+0.01085$ at full disclosure $[0.00444, 0.01741]$. Intervals resample complete paired simulation replicates and quantify Monte Carlo uncertainty under this specific generator and target. They do not establish real-world superiority.

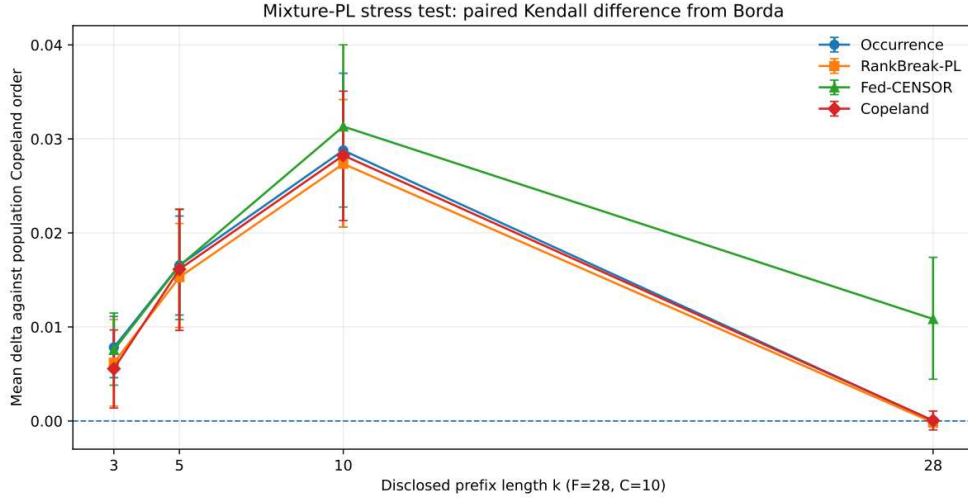


Figure 6: Paired Fed-CENSOR-minus-Borda Kendall differences for the specified two-component mixture-PL generator, with percentile bootstrap intervals over 100 replicates per disclosure level.

9.4 Censoring-isolation and random-tail completion controls

To isolate the effect of respecting the unobserved tail, an independent common-PL control uses 50 replicates per disclosure level. The same prefix is fitted either by censored PL-MLE or by full-ranking PL-MLE after deterministic feature-index completion of the hidden tail. Relative to the known generating-worth ranking, censored PL-MLE obtains Kendall $\tau = 0.466667$ versus 0.006138 at $k = 3$ (difference 0.460529), 0.568254 versus 0.042011 at $k = 5$ (difference 0.526243), and 0.688995 versus 0.226455 at $k = 10$ (difference 0.462540). At $k = 28$ the procedures coincide at 0.810688 . This synthetic control shows that fabricating a tail order can distort PL estimation under a known PL generator; it is not a competitive real-data baseline.

A second control randomizes the hidden tail before full-ranking PL fitting. The retained true-ranking Kendall means (Borda, censored PL, random-tail PL) are $(0.443624, 0.453571,$

0.183471) at $k = 3$, (0.540582, 0.555265, 0.305053) at $k = 5$, and (0.655873, 0.680026, 0.518270) at $k = 10$. Censored-versus-random-tail differences are 0.270101, 0.250212, and 0.161757, respectively. This is an imputation stress test, not a competitive aggregation baseline.

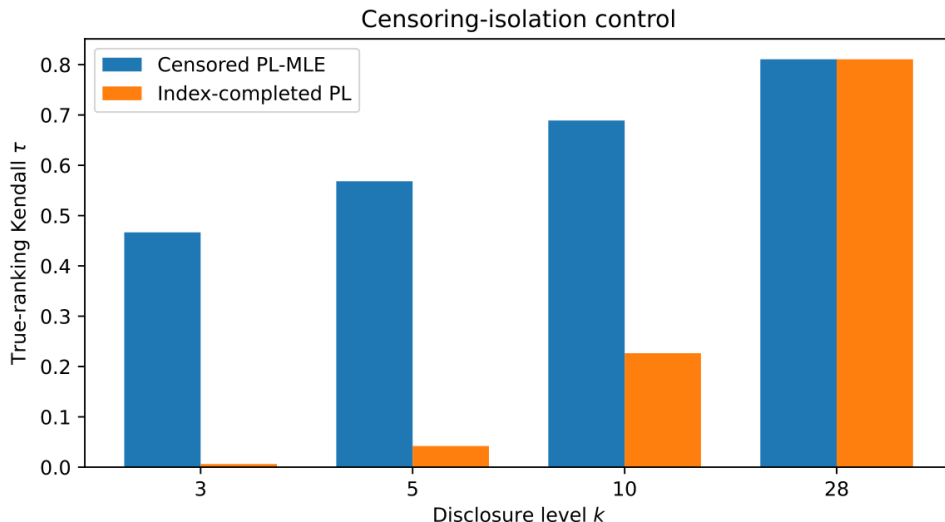


Figure 7: Censoring-isolation control comparing censored PL-MLE with PL-MLE fitted after deterministic completion of the hidden tail. The result is a synthetic diagnostic, not a real-benchmark advantage.

9.5 Full-ranking/no-censoring ablation

An additional ten-seed ablation (seeds 121–130) compares unregularized PL MLE with Borda on the same benchmark under masked and full-ranking inputs. Here, masked means only the ordered top- k prefix is supplied and no within-tail order is observed; full-ranking means the complete client permutation is supplied. For masked inputs, the seed-averaged PL-MLE-minus-Borda difference is -0.00212 at $k = 3$ ($p = 0.2754$) and -0.00437 at $k = 10$ ($p = 0.3750$). Under full information, the nominal $k = 3, 10, 28$ labels refer to the same complete client permutations and are not independent conditions. After averaging the four model-specific differences within each seed, all ten full-ranking seed-block differences are negative, with mean -0.035714 Kendall points and exact two-sided signed-rank $p = 0.001953$. The ten values are reproduced in Appendix Table A4. This is a control against the claim that unregularized PL estimation is automatically preferable to Borda under complete rankings.

9.6 Same-prefix baseline suite

The additional baseline suite uses the same retained prefixes and centralized-reference rankings. Re-derived Weighted Borda and Fed-CENSOR Kendall scores matched stored per-job values for all 80 seed-model- k comparisons at $k \in \{3, 10\}$, with maximum absolute discrepancy below 10^{-10} . The available weighting helper also matched the reconstructed training-count calculation for 40/40 seed-model checks. Because the original training runner

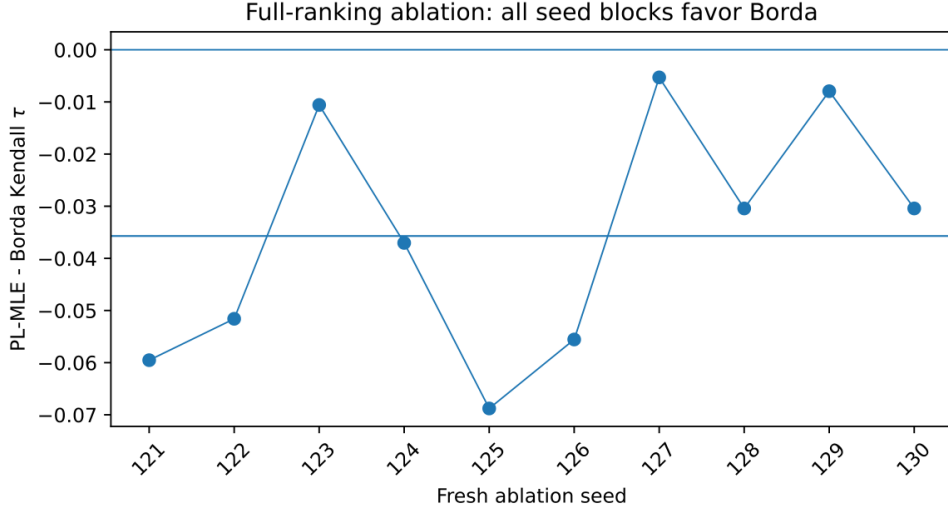


Figure 8: Ten-seed full-ranking ablation. Inference uses one seed block per seed after averaging the four model-specific differences; nominal disclosure labels under full information are not separate replicates.

is absent, this is a consistency audit, not an independent source-level reproduction of the original weighting implementation.

Table 7: Same-prefix baseline suite. Mean centralized-reference Kendall across retained seed blocks and model families. Pairwise methods are not identifiable at $k = 1$, where no within-prefix feature pair is observed.

k	Weighted Borda	Occurrence count	Pairwise BT	Fed-CENSOR	Copeland margin
1	0.364683	0.364683	N/I	0.364683	N/I
3	0.404101	0.403571	0.386640	0.404101	0.177910
5	0.444974	0.449206	0.400926	0.447354	0.131878
10	0.521825	0.506349	0.458201	0.516667	0.212434

The Pairwise BT column in Table 7 evaluates a strict consensus permutation after stable tie-breaking; Table A13’s regularized-BT full-vector analysis instead computes Kendall τ_b directly on score vectors and retains ties. These results are not directly comparable effect estimates. In the regularized-BT representation, large tied/weakly identified feature groups affect τ_b , especially outside the graph of observed pairwise comparisons. The descriptive occurrence-count means at $k = 3, 5, 10$ are 0.403571, 0.449206, and 0.506349, compared with 0.404101, 0.444974, and 0.521825 for Borda. These are averages of repeated rows, not independent replicates. The retained-prefix tail-score sweep shows that Borda’s convention matters: across four unoptimized tail assignments, mean Kendall ranges are 0.392–0.404 at $k = 3$, 0.424–0.448 at $k = 5$, and 0.513–0.522 at $k = 10$; mean NDCG@10 ranges are 0.667–0.679, 0.693–0.704, and 0.649–0.688, respectively (Appendix Table A15). These descriptive ranges do not select a best convention and show why the default truncated-Borda rule should not be treated as uniquely determined by the observation model. The older pairwise rank-breaking PL-MM implementation converged in only 7/40, 12/40, and 21/40 fits at

$k = 3, 5, 10$ (17.5%, 30.0%, 52.5%); non-converged outputs are not treated as valid rankings. The regularized Bradley–Terry LBFGS audit is separate: all 120 optimizer runs terminated successfully, 53/120 met the requested gradient tolerance of 10^{-9} , and 120/120 met the review threshold of 10^{-7} . Pairwise BT is not identifiable at $k = 1$. The Copeland-margin diagnostic uses only co-disclosed pairs and is an exploratory comparator, not a state-of-the-art baseline.

A separate tie-aware exploratory analysis tested 18 planned exploratory contrasts: three alternative methods versus Borda, three disclosure levels, and two metrics (full-vector Kendall τ_b and co-disclosed-pair Kendall). Four contrasts survived Holm correction, all involving regularized Bradley–Terry: the full-vector metric at $k = 3, 5, 10$ and co-disclosed-pair Kendall at $k = 10$. All four mean differences were negative. Given the narrow co-disclosed-pair mask and the fact that full-vector score geometry includes unresolved tail order, these exploratory results do not establish general superiority of Borda. All numeric contrasts and confidence intervals are in Appendix Table A13.

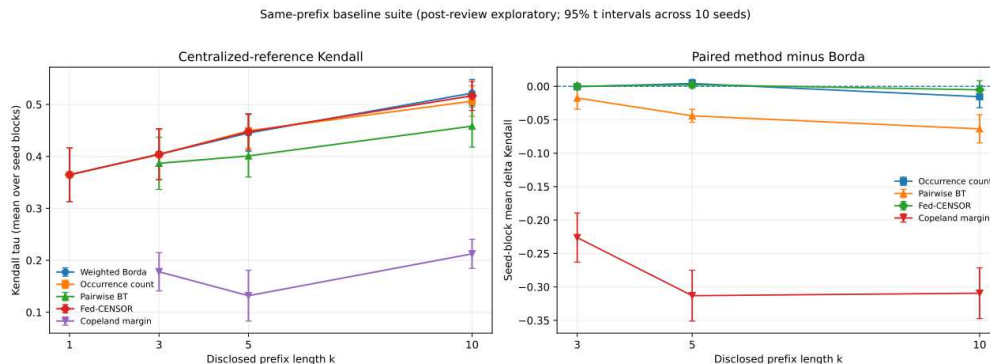


Figure 9: Same-prefix baseline suite. This descriptive figure is not a substitute for seed-block inference and should be interpreted alongside the convergence and pair-coverage limitations.

The suite does not cover every established baseline family. Spectral/Luce optimization, Kemeny or footrule optimization, value averaging, and a fully specified MAP-PL comparator remain unexecuted in the supplied artifact bundle. The current suite is not comprehensive, and no universal superiority claim is made.

For artifact concordance, the direct confirmatory per-job table gives mean Fed-CENSOR Kendall agreement of 0.404101 at $k = 3$ and 0.516667 at $k = 10$, while the separately derived nested-disclosure summary gives 0.403968 and 0.516402. The Borda means agree to displayed precision, but the Fed-CENSOR means differ slightly. The primary result uses directly matched confirmatory per-job rows; nested-disclosure values remain a separate derived artifact and are not substituted for the primary endpoint.

10 Heterogeneity and Model-Family Analysis

The historical factorial study varies federation size across 3, 5, and 8 clients and Dirichlet alpha across 0.1, 0.5, and 1.0. Descriptively, mean Spearman agreement is approximately 0.593 at alpha=0.1 and 0.617 at alpha=1.0, while the corresponding means for 3 and 8 clients are approximately 0.617 and 0.594. These summaries should not be read as causal

estimates because the independent unit is the seed and the factorial design contains repeated configuration measurements within seed.

Model-family differences are also material. Logistic regression, MLP, Random Forest, and XGBoost differ in both predictive behavior and the federation mechanism used by the implementation. RF/XGBoost use weighted probability ensembles rather than parameter FedAvg, so model-family comparisons are descriptive stress tests rather than a clean four-model causal comparison of one FL algorithm. The realized client positive-class proportion across confirmatory seeds spans approximately 0.008 to 0.988 (mean 0.551, SD 0.337), confirming substantial residual label skew after repair.

11 Ablations and Sensitivity

11.1 Weighting sensitivity

The retained sensitivity summary compares sample-size weighting with uniform weighting and reports aggregation-reference scores. These descriptive values do not replace the confirmatory comparison.

Table 8. Weighting sensitivity. Retained summaries for $k = 3$ and $k = 10$; not a confirmatory test family.

	Borda vs. k	PL vs. uniform ρ	Uniform Borda/PL ρ/τ	Borda/Fed-C. τ
3	0.9946	0.9938	0.9983 / 0.9901	0.4041 / 0.4040
10	0.9666	0.9525	0.9844 / 0.9402	0.5218 / 0.5164

12 External Dataset Validation

The external-validation branch tests whether the observed aggregation behavior transfers beyond the pooled UCI Heart Disease benchmark. Two protocols must be kept distinct: the retained one-shot run (seed 301) uses four model families and nominal k values $\{1, 3, 5, 10\}$; the ten-seed WDBC extension uses ten seed blocks and a single RandomForest-TreeSHAP ranking pipeline. The repeated-seed extension supplements the one-shot rows, and neither changes the primary Heart Disease endpoint.

Dataset summary.

Dataset	Rows	Features	Positive	Negative	Nominal k
Breast Cancer Wisconsin (Diagnostic)	569	30	212	357	1, 3, 5, 10
Pima Indians Diabetes	768	8	268	500	1, 3, 5, 10*

*For Diabetes, nominal $k = 10$ is effectively full disclosure because the dataset has only eight predictors.

12.1 Breast Cancer Wisconsin Diagnostic: ten-seed extension

The ten-seed extension uses seeds 101-110, five simulated clients, 30 numeric predictors, sample-size weighting, local RandomForest models, and TreeSHAP version 0.52.0. The

aggregation run uses local explanation budget 20 and reports 90 rows: ten seeds by three disclosure levels ($k = 3, 5, 10$) by three methods. The centralized reference is evaluated on 100 explanation rows. Across the ten seed-specific centralized reference models, mean accuracy is 0.9531 (SD 0.0162) and mean ROC-AUC is 0.9890 (SD 0.0087); these figures describe this dataset protocol, not clinical performance.

12.2 Single-seed WDBC analysis

The seed-301 analysis contains 32 method/model/ k rows. At $k = 1$ the two main aggregators were identical. Its mean Kendall differences (Fed-CENSOR minus Borda) were +0.0030 at $k = 3$, +0.0113 at $k = 5$, and +0.0080 at $k = 10$, while Spearman differences changed sign. Because that branch has one seed, its results remain descriptive and are not pooled with the new ten-seed extension.

Ten-seed WDBC aggregation results. Entries are mean Kendall tau-b (SD) against the centralized reference; overlap is top-10 overlap under the stable tie-break.

k	Method	Kendall tau-b, mean (SD)	Top-10 overlap
3	Weighted Borda	0.5904 (0.0304)	0.81
3	Weighted occurrence	0.5890 (0.0327)	0.81
3	Censored PL-MM	0.5897 (0.0331)	0.81
5	Weighted Borda	0.6678 (0.0252)	0.89
5	Weighted occurrence	0.6777 (0.0319)	0.89
5	Censored PL-MM	0.6664 (0.0261)	0.89
10	Weighted Borda	0.7550 (0.0345)	0.92
10	Weighted occurrence	0.7476 (0.0384)	0.89
10	Censored PL-MM	0.7546 (0.0343)	0.92

All six seed-block comparisons of occurrence count and censored PL-MM against Borda (two contrasts at each of three k values) were non-significant after Holm correction. Weighted occurrence at $k = 5$ had unadjusted exact $p = 0.048828$, but adjusted $p = 0.292969$. The repeated-seed extension therefore gives no multiplicity-adjusted evidence that any method is better on WDBC. Tie-aware overlap bounds and the remaining per-seed rows are retained in the machine-readable output.

12.3 Diabetes: retained one-shot analysis

The Pima Indians Diabetes dataset has 768 observations and eight predictors (Smith et al., 1988). Clinically implausible zeros in glucose, blood pressure, skin thickness, insulin, and BMI are treated as missing before training-partition median imputation and standardization. Five repaired-Dirichlet clients are used in the one-shot seed-301 branch. Mean Kendall agreement is -0.0714 for both methods at nominal $k = 1$; Borda versus Fed-CENSOR is 0.3036 versus 0.3750 at $k = 3$, 0.3929 versus 0.5000 at $k = 5$, and 0.5179 versus 0.5893 at nominal $k = 10$. Spearman differences are +0.1071, +0.0774, and +0.0655 at $k = 3, 5$, and 10. These are single-seed descriptive contrasts; nominal $k = 10$ is full disclosure for this eight-feature dataset.

12.4 Interpretation

The repeated-seed WDBC extension is more informative than the earlier one-shot run about seed-to-seed variability, but it remains an exploratory tabular experiment using simulated clients and a model-derived reference. It does not establish generalization to other modalities, sites, or explanation algorithms. The Diabetes results remain single-seed and cannot support replicated superiority claims.

The one-shot Diabetes contrasts are descriptive and should not be interpreted as evidence that censoring itself produced a gain. No external statistical test is treated as confirmatory because that branch uses one seed and repeated model families are not independent replication units.

13 Statistical Analysis and Power

The independent inferential unit is the seed block after averaging paired differences across the four model families. The five-seed historical factorial study has coarse exact-test resolution and is reported descriptively; the ten-seed confirmatory analysis is the primary inferential study. The confirmatory primary endpoint is the mean seed-level difference in centralized-reference Kendall agreement at $k = 10$. A tie-aware exact signed-rank sensitivity uses average midranks for tied absolute differences and enumerates all 2^{10} sign assignments.

The 11 secondary contrasts comprise six metrics at $k = 3$ and five non-primary metrics at $k = 10$. Holm correction is applied to that family; the $k = 10$ primary endpoint is not included in the secondary family. NDCG@10 favors Borda and remains significant after correction, but the endpoint was not prospectively registered. A non-significant primary test is not evidence of equivalence: no scientifically justified smallest effect size of interest was specified in advance, so no formal TOST or ROPE claim is made. Some $k = 3$ set-metric values and the tie-aware score-vector sensitivity are recomputed from retained client ranking payloads and reconstructed weights; they are not independent training or SHAP reruns.

14 Discussion

The central result is deliberately modest. On the retained real benchmark, Fed-CENSOR does not establish a performance improvement over sample-size-weighted Borda under the primary centralized-reference Kendall endpoint. The ten-seed mean difference at $k = 10$ is -0.00516 , with tie-aware exact signed-rank $p = 0.453$ and a 95% t interval $[-0.01873, 0.00841]$; five seed blocks favor each method. The retained 20-seed extension includes seeds 101–110 as well as 111–120; row-level output for the latter subset is not available in the supplied package, so we do not report it as a separately verified independent replication.

The prefix-resolved analysis explicitly includes disclosed-versus-omitted relations. Using all 378 unordered pairs as the denominator, total prefix-resolved coverage is 45.9%, 61.2%, and 87.2% at $k = 3, 5, 10$, respectively. The narrower co-disclosed-pair statistic is separately labeled in Appendix Table A13. The co-disclosed-pair sensitivity does not reveal a reliable Fed-CENSOR advantage. NDCG@10 favors Borda in all ten seed blocks and remains significant after Holm correction over 11 secondary contrasts; it should be followed up because it directly emphasizes top-feature recovery, while remaining a secondary non-prospectively registered endpoint.

Synthetic evidence shows a conditional use case. Under a known common-PL generator, the method difference changes with disclosure length, and the censoring-isolation control shows that fabricated tail order can distort PL estimation. Those experiments support respecting the observation model; they do not establish that real SHAP rankings follow a common PL distribution closely enough to yield a benchmark advantage. The mixture-PL experiment uses a particular, fully specified generator and population Copeland target, so its positive deltas remain generator-specific.

The full-ranking ablation provides a negative control: in the seeds 121–130 ablation, unregularized full-ranking PL-MLE is below Borda by 0.035714 Kendall points (exact signed-rank $p = 0.001953$). This does not replace the original primary study because the seeds and full-ranking protocol differ. The broader same-prefix suite adds occurrence count, pairwise BT, and Copeland diagnostics; low rank-breaking convergence and the narrower co-disclosed-pair mask limit some exploratory comparisons, while the regularized BT-LBFGS audit is kept separate because its estimator, optimization status, and tie-aware metric differ.

The new explanation-budget study addresses one specific reproducibility gap: in WDBC, TreeSHAP rankings are highly but not perfectly stable across 5, 10, and 20 explanation examples (mean Kendall 0.907-0.947; mean top-10 overlap 0.972-0.980). The ten-seed WDBC aggregation study shows similar reference agreement for Borda and censored PL-MM, with no adjusted significant method contrast. These are useful sensitivity results, not evidence that the main Heart Disease SHAP pipeline is budget-invariant: the latter uses LinearExplainer, TreeExplainer, and KernelSHAP across different model families, and the WDBC study covers only RandomForest-TreeSHAP.

The comparative conclusion is therefore conservative. Weighted Borda remains the simpler practical default supported by the current evidence in this benchmark; the study does not establish universal superiority. Fed-CENSOR is a model-based alternative when a common PL structure is plausible, disclosure supplies enough information, and additional fitting cost is acceptable. The present study does not establish how often those conditions hold in real cross-silo deployments.

The paper also has methodological limits: the four model families use different federation mechanisms, so model-family comparisons are descriptive rather than causal; the main Heart Disease explanation budget is small; and the original training/SHAP runner is absent.

15 Limitations

1. Benchmark and deployment scope. The main federated study uses one public tabular benchmark with 920 source records. External WDBC and Diabetes analyses are secondary checks rather than full, independently replicated multi-round FL studies.

2. Simulated clients. Clients are produced by constrained/repaired Dirichlet partitioning. The natural UCI source cohorts are source domains or cohort simulations, not real hospitals; no clinical or prospective validation is claimed.

3. Reference-model uncertainty. The centralized reference is a separately trained, model-derived ranking, not ground truth. Agreement with it is not absolute feature-importance correctness.

4. Model-family confounding. LR and MLP use parameter FedAvg, whereas RF and XGBoost use local retraining with sample-size-weighted probability ensembling. Model-family comparisons therefore mix model and federation-mechanism effects.

5. Data coding and feature grouping. Median/mode imputation and one-hot expansion are applied to the supplied Heart Disease table, but zero-coded clinical fields have not been audited as possible missingness indicators separately in each source cohort. One-hot columns are ranked individually rather than aggregated back to their originating categorical variable. These choices can affect local explanation rankings and require a dedicated preprocessing/group-level sensitivity analysis.

6. Explanation budget and predictive configuration. The confirmatory Heart Disease run uses five background samples, three client explanation samples, and 32 KernelSHAP samples for the MLP path; this small explanation budget may produce noisy local rankings. The WDBC TreeSHAP budget experiment uses a different dataset and explainer and does not answer whether the Heart Disease result is stable at larger budgets. A matched higher-budget Heart Disease rerun has not been executed. The XGBoost configuration is conservative and was not tuned for predictive performance, so model-family summaries are not a tuned predictive benchmark. A separate WDBC experiment now tests TreeSHAP budgets 5, 10, and 20 over ten seeds and five clients, and finds high but imperfect within-client stability. It does not establish budget invariance for other explainers or datasets.

7. Baseline coverage. The comparison suite includes weighted Borda, occurrence count, rank-breaking PL, regularized BT, and Copeland diagnostics. In the convergence audit, RankBreak-PL-MM converged in 17.5%, 30.0%, and 52.5% of fits at $k = 3, 5$, and 10; non-converged fits are excluded from valid-score inference. Regularized BT-LBFGS converged in 40/40 runs for each of those k values, but $k = 1$ is not identifiable from pairwise comparisons. A separate pairwise-BT/Copeland additional suite has its own metric and coverage limitations and is not conflated with this convergence audit. Luce spectral/ILSR, MAP or Bayesian PL with an explicit prior, an average-tied-tail completion control, an independent cross-check against the R PlackettLuce implementation, and a same-aggregator full-ranking reference comparison remain unexecuted. These are priority follow-ups, not negative results.

8. Synthetic stress calibration. The common-PL generator is fully specified, but exact perturbation magnitudes for retained client-heterogeneity and rank-noise stress regimes are not recoverable. Those results are descriptive rather than calibrated generative inference.

9. Numerical and computational scope. The MM stopping rule checks parameter change but does not independently verify comparison-graph connectivity or likelihood improvement. Runtime results measure aggregation routines on retained payloads and exclude training, SHAP computation, I/O, transport, retries, and network conditions.

10. Statistical power. The primary confirmatory study uses ten independent seed blocks; the historical study uses five. No equivalence margin was prespecified, so a non-significant result is not interpreted as equivalence.

16 Reproducibility and Artifact

The artifact package contains retained primary-result tables, baseline diagnostics, seed-level summaries, WDBC explanation-budget scores and stability summaries, aggregation outputs, and input validation. These retained analyses do not include the original end-to-end model-

training and SHAP runner. The 4,500 SHAP score rows cover ten seeds, five clients, 30 features, and budgets 5, 10, and 20; the 90 WDBC aggregation rows cover ten seeds, three k values, and three methods.

The confirmatory primary endpoint is reported separately from all derived sensitivity analyses. The independent source-code cross-check for reconstructed sample-size weights remains open because the original training runner is not present in the supplied artifact bundle. The retained partition files match stored size metadata and the available weighting-helper checks pass, but these checks are not a substitute for a clean reproduction from the original source. An executable public release should include the original runner, environment lockfile, raw-input checksums, and persistent versioned artifacts. The remaining unexecuted comparators are listed explicitly rather than treated as negative results.

The historical five-seed and ten-seed confirmatory studies remain the main inferential evidence. The SHAP-budget and ten-seed WDBC analyses are explicitly exploratory. The available 20-seed extension includes seeds 101–110; the row-level results for seeds 111–120 are not present in the supplied package, so those ten seeds cannot be independently summarized here. The study was not prospectively registered. The primary endpoint is reported as the confirmatory analysis endpoint, not as a prospectively registered hypothesis.

17 Conclusion

This paper evaluates an established prefix-censored Plackett–Luce likelihood for aggregating top- k explanation rankings in a simulated federated setting. It does not introduce a new likelihood or MM estimator. On the UCI Heart Disease benchmark, the primary-analysis $k = 10$ Kendall endpoint does not establish improved agreement over sample-size-weighted Borda: the mean difference is -0.00516 , the tie-aware exact signed-rank p -value is 0.453 , the 95% t interval is $[-0.01873, 0.00841]$, and five seed blocks favor each method.

The secondary NDCG@10 contrast favors Borda and remains significant after correction over 11 secondary contrasts; because it was not prospectively registered, it is evidence to follow up rather than a stand-alone basis for a superiority claim. The corrected prefix-resolved analysis shows that the top- k observation carries more pairwise information than a within-prefix-only mask suggests: coverage rises from 45.9% at $k = 3$ to 87.2% at $k = 10$. At the same time, a large number of features remain undisclosed at small k , and the results depend in part on the Borda tail-score convention and tie handling.

The controlled synthetic results show that censored PL can recover rankings better under some generators and disclosure levels, but not uniformly. The real-benchmark results, the full-ranking ablation, and the tie-aware sensitivity do not establish a general advantage for PL aggregation. The highest-priority next steps are a matched higher-budget Heart Disease SHAP study, stronger partial-ranking comparators (including MAP/Bayesian and spectral PL methods), a cross-check against a maintained PL implementation, and end-to-end reproduction from the original training and explanation-generation runner. Until those gaps are closed, the appropriate conclusion is an empirical characterization of when the censoring-aware likelihood changes the result, not a claim of universal superiority.

A Supplementary Evidence and Analysis Inventory

The following analyses are separate from the primary confirmatory test and are interpreted as supplementary validation or sensitivity analyses.

Supplementary evidence inventory.

Experiment	Status	Interpretation
Synthetic PL ground truth	Executed	1,200 rows across three regimes, four disclosure levels, and 100 replicates per condition.
Censoring-isolation control	Executed; supplementary	50 common-PL replicates per disclosure level; deterministic-tail completion control.
Prefix-resolved pairs	Recomputed from retained payloads	45.9%, 61.2%, and 87.2% pair coverage at k=3, 5, and 10; k=5 is nested from k=10.
Full-ranking ablation	Executed	Ten fresh seeds; unmasked PL below Borda by 0.0357 Kendall.
Same-prefix baseline suite	Executed; exploratory	Occurrence, pairwise BT, and Copeland; convergence/coverage limits remain.
Mixture-PL stress test	Executed; supplementary	400 paired replicates; target is the population Copeland order for this generator.
SHAP budget sensitivity	Executed; exploratory	4,500 WDBC TreeSHAP scores; budgets 5/10/20; 50 seed-client comparisons per budget pair.
WDBC repeated-seed aggregation	Executed; exploratory	90 method/seed/k rows; no paired contrast significant after Holm correction.
Realized client heterogeneity	Executed	Positive-class proportions 0.008–0.988 across confirmatory clients.
Optimized runtime	Executed	Cumulative-sum implementation reduces aggregation routine time by 2.40x–3.59x.
Source-domain hold-out	Not reported	Row-level quantitative outputs are unavailable in the supplied package.
Matched Heart Disease SHAP budget	Not executed	Main three-example explanation budget has not been rerun at larger budgets.
MAP/Bayesian and spectral PL baselines	Not executed	No result is implied by their absence.
R PlackettLuce cross-check	Not executed	The original end-to-end runner is missing.
Formal equivalence test	Not executed	No independently justified smallest effect size of interest was established.
End-to-end network measurement	Not executed	No packet-level or transport benchmark is claimed.
Real cross-silo validation	Not executed	Clients remain simulated.

B Exact primary and extended result tables

For the client-heterogeneity and rank-noise rows, the retained exact p values are reported as computational summaries only; the corresponding perturbation magnitudes are not fully recoverable, so these rows are not used for calibrated inferential claims.

The random-tail fit is an imputation stress test, not a competitive aggregation baseline.

Table A1. Primary endpoint.

Quantity	Result
Independent seed blocks	10
Mean Fed-CENSOR minus Borda Kendall τ at $k = 10$	-0.005159
Tie-aware exact signed-rank p	0.453125
Cohen’s paired d_z	-0.272
95% t interval for mean difference	$[-0.01873,$ 0.00841]
Seed blocks favoring Fed-CENSOR / Borda	5 / 5

Table A2. Frozen 20-seed extension.

Quantity	Value
Seeds	101–120
Mean Delta Kendall@10	-0.0049603174603174
Bootstrap 95% CI	$[-0.0113756613756613, 0.0013888888888889]$
Exact Wilcoxon p	0.2454872131347656
Positive / negative / zero	8 / 12 / 0

Table A3. Full-ranking ablation.

Information	k	Borda tau	Unregularized PL-MLE tau	Mean Delta	Exact p
Masked	3	0.430159	0.428042	-0.002116	0.275391
Masked	10	0.522751	0.518386	-0.004365	0.375000
Full	3	0.555556	0.519841	-0.035714	0.001953
Full	10	0.555556	0.519841	-0.035714	0.001953
Full	28	0.555556	0.519841	-0.035714	0.001953

Table A4. Full-ranking seed-block differences. Values are seed blocks 121–130.

Seed	PL MLE minus Borda Kendall difference
121	-0.059524
122	-0.051587
123	-0.010582
124	-0.037037
125	-0.068783
126	-0.055556
127	-0.005291
128	-0.030423
129	-0.007937
130	-0.030423
Mean	-0.035714
Exact two-sided signed-rank p	0.001953

Table A5. Censoring-isolation control summary.

k	Censored PL-MLE τ	Index-completed PL τ	Mean difference
3	0.466667	0.006138	0.460529
5	0.568254	0.042011	0.526243
10	0.688995	0.226455	0.462540
28	0.810688	0.810688	0.000000

Table A6. Ground-truth synthetic validation. Mean Fed-CENSOR-minus-Borda Kendall differences over 100 simulation replicates per regime and disclosure level, paired bootstrap 95% intervals, and fraction of replicates with a positive difference. We emphasize effect sizes and uncertainty under each fixed generator rather than very small Monte Carlo p-values.

Regime	k	Mean $\Delta\tau$	Paired bootstrap 95% interval	Positive fraction
Common PL	3	-0.01222	[-0.01693, -0.00725]	0.22
Common PL	5	-0.00831	[-0.01360, -0.00302]	0.31
Common PL	10	+0.01767	[+0.01116, +0.02439]	0.70
Common PL	28	+0.02720	[+0.02164, +0.03286]	0.80
Client heterogeneity	3	-0.01286	[-0.01720, -0.00862]	0.23
Client heterogeneity	5	-0.00550	[-0.01175, +0.00090]	0.33
Client heterogeneity	10	+0.01138	[+0.00392, +0.01889]	0.56
Client heterogeneity	28	-0.00185	[-0.00884, +0.00503]	0.43
Rank noise	3	-0.01111	[-0.01529, -0.00693]	0.23
Rank noise	5	-0.00063	[-0.00698, +0.00571]	0.46
Rank noise	10	+0.02143	[+0.01428, +0.02894]	0.68
Rank noise	28	+0.01968	[+0.01296, +0.02646]	0.70

Table A7. Random-tail completion stress test.

k	Borda τ	Censored PL τ	Random-tail PL τ	Censored PL – Random-tail PL
3	0.443624	0.453571	0.183471	0.270101
5	0.540582	0.555265	0.305053	0.250212
10	0.655873	0.680026	0.518270	0.161757

Table A8. Baseline coverage and status. Status reflects the evidence available in the released artifact; unexecuted baselines are not interpreted as negative results.

Item	Status
Weighted Borda	Evaluated on the same disclosed client prefixes as censored PL-MM.
Weighted occurrence count	Evaluated as a simple non-parametric supplementary baseline.
Same-prefix baseline suite	80 stored seed/model/ k score comparisons were checked. Pair-wise BT is not identifiable at $k = 1$; Copeland uses only co-disclosed pairs.
Partitioned-preference PL relation	The top- k prefix is a special case of ordered-partition Plackett-Luce observations (Ma et al., 2021); an independent cross-implementation check of the estimator is not available in this package.
Spectral/Luce, Kemeny/footrule, value averaging, MAP/Bayesian PL, average-tied-tail PL completion	Not executed in the supplied artifact.
Same-aggregator full-ranking reference comparison	Not executed; needed to isolate censoring loss from centralized-reference error.

Table A9. Prefix-resolved pair analysis. Full-vector τ_b means are a convergence-gated matched sensitivity and match Table A14; the matched rows have 39/40 valid PL fits at each k (seed 109 has three matched model families). Prefix-resolved Kendall uses the same matched rows and only pairs with at least one feature disclosed by any client. Pair coverage is computed across all 40 input blocks, independent of estimator convergence.

k	Method	Valid rows	Full-vector τ_b	Prefix-resolved τ	Coverage
3	Weighted Borda	39/40	0.44326	0.66790	45.93%
3	Censored PL-MM	39/40	0.44336	0.66805	45.93%
5	Weighted Borda	39/40	0.49076	0.63856	61.22%
5	Censored PL-MM	39/40	0.49380	0.64104	61.22%
10	Weighted Borda	39/40	0.53147	0.57266	87.19%
10	Censored PL-MM	39/40	0.52624	0.56596	87.19%

Table A10. Optimized aggregation runtime.

k	Borda ms	Fed-CENSOR optimized ms	Ratio
3	0.016520	0.249374	15.10
10	0.016537	0.876702	53.01

Table A11. Exploratory same-prefix seed-block contrasts against Weighted Borda. Each difference is averaged over the four model families within each seed; the family of 12 method-by- k contrasts is Holm-adjusted. Pairwise BT and Copeland rows use the retained score representations and should not be conflated with the PLMM convergence audit.

Method	Metric	k	n	Mean Δ	95% CI	Exact p	Holm p
Fed-CENSOR	Kendall	3	10	+0.000000	[-0.0027, 0.0027]	1.000000	1.000000
Fed-CENSOR	Kendall	5	10	+0.002381	[-0.0025, 0.0073]	0.322266	1.000000
Fed-CENSOR	Kendall	10	10	-0.005159	[-0.0187, 0.0084]	0.453125	1.000000
Pairwise BT	Kendall	3	10	-0.017460	[-0.0343, -0.0006]	0.048828	0.341797
Pairwise BT	Kendall	5	10	-0.044048	[-0.0537, -0.0344]	0.001953	0.023438
Pairwise BT	Kendall	10	10	-0.063624	[-0.0848, -0.0425]	0.001953	0.023438
Copeland margin	Kendall	3	10	-0.226190	[-0.2630, -0.1893]	0.001953	0.023438
Copeland margin	Kendall	5	10	-0.313095	[-0.3511, -0.2751]	0.001953	0.023438
Copeland margin	Kendall	10	10	-0.309392	[-0.3473, -0.2715]	0.001953	0.023438
Occurrence count	Kendall	3	10	-0.000529	[-0.0032, 0.0021]	0.820312	1.000000
Occurrence count	Kendall	5	10	+0.004233	[-0.0020, 0.0104]	0.160156	0.800781
Occurrence count	Kendall	10	10	-0.015476	[-0.0321, 0.0011]	0.130859	0.785156

Table A12. Synthetic mixture-PL paired difference from Borda. Bootstrap intervals resample complete paired simulation replicates (5,000 bootstrap resamples). These intervals quantify Monte Carlo uncertainty for the fixed generator and target, not uncertainty for a real-world population.

k	Fed-CENSOR mean Δ	Bootstrap 95% interval	Occurrence mean Δ	Copeland mean Δ
3	+0.007566	[+0.003810, +0.011481]	+0.007831	+0.005556
5	+0.016508	[+0.010792, +0.022487]	+0.016561	+0.016138
10	+0.031323	[+0.022750, +0.040000]	+0.028783	+0.028254
28	+0.010847	[+0.004444, +0.017409]	+0.000000	+0.000053

Table A13. Tie-aware baseline contrasts versus Borda. Differences are averaged across four model families within each seed. The Holm family includes all 18 planned exploratory contrasts; negative differences favor Borda on the metric shown. The co-disclosed-pair rows use only pairs appearing together within one disclosed prefix.

Metric	Alternative method	k	n	Mean Δ	95% CI	Exact p	Holm p
Full-vector τ_b	Occurrence	3	10	+0.001089	[-0.003464, 0.005642]	0.625000	1.000000
Full-vector τ_b	Censored PL-MM	3	9	-0.000714	[-0.004785, 0.003357]	0.652344	1.000000
Full-vector τ_b	Regularized BT	3	10	-0.420059	[-0.488955, -0.351163]	0.001953	0.035156
Full-vector τ_b	Occurrence	5	10	+0.006371	[-0.001815, 0.014557]	0.105469	1.000000
Full-vector τ_b	Censored PL-MM	5	9	+0.000760	[-0.004996, 0.006515]	0.734375	1.000000
Full-vector τ_b	Regularized BT	5	10	-0.412191	[-0.460761, -0.363621]	0.001953	0.035156
Full-vector τ_b	Occurrence	10	10	-0.015206	[-0.031696, 0.001284]	0.105469	1.000000
Full-vector τ_b	Censored PL-MM	10	9	-0.007932	[-0.023111, 0.007248]	0.250000	1.000000
Full-vector τ_b	Regularized BT	10	10	-0.299269	[-0.349396, -0.249142]	0.001953	0.035156
Co-disclosed-pair τ	Occurrence	3	10	+0.052340	[-0.028384, 0.133065]	0.160156	1.000000
Co-disclosed-pair τ	Censored PL-MM	3	9	+0.000661	[-0.039441, 0.040764]	0.945313	1.000000
Co-disclosed-pair τ	Regularized BT	3	10	-0.105441	[-0.208535, -0.002347]	0.083984	1.000000
Co-disclosed-pair τ	Occurrence	5	10	+0.034258	[-0.016883, 0.085398]	0.193359	1.000000
Co-disclosed-pair τ	Censored PL-MM	5	9	-0.006710	[-0.036907, 0.023486]	0.652344	1.000000
Co-disclosed-pair τ	Regularized BT	5	10	-0.157563	[-0.269493, -0.045633]	0.019531	0.273438
Co-disclosed-pair τ	Occurrence	10	10	-0.014850	[-0.057412, 0.027713]	0.492188	1.000000
Co-disclosed-pair τ	Censored PL-MM	10	9	-0.016777	[-0.040720, 0.007166]	0.425781	1.000000
Co-disclosed-pair τ	Regularized BT	10	10	-0.111631	[-0.143465, -0.079798]	0.001953	0.035156

Four contrasts survive Holm correction: regularized BT on full-vector τ_b at $k = 3, 5, 10$ and on co-disclosed-pair τ at $k = 10$. The latter metric includes only pairs that co-occur within a disclosed prefix, not all prefix-versus-tail relations. These are exploratory contrasts.

Table A14. Convergence-gated tie-aware score-level Kendall sensitivity. Kendall’s τ_b is computed on Borda score and PL worth vectors, preserving score ties instead of imposing feature-index order. Only matched seed/model rows whose extended PL fit passes the 5,000-iteration / 10^{-8} stopping criterion are included; 39/40 rows pass at each k , and seed 109 contributes three model families. $k = 5$ is nested from $k = 10$.

k	Borda mean τ_b	PL mean τ_b	Mean PL–Borda	Exact p	95% t interval
3	0.44326	0.44336	+0.00010	1.000000	[-0.00392, 0.00412]
5	0.49076	0.49380	+0.00304	0.556641	[-0.00441, 0.01049]
10	0.53147	0.52624	-0.00522	0.492188	[-0.01989, 0.00944]

Exact two-sided signed-rank p values enumerate all 2^{10} sign assignments using average midranks for tied absolute differences. The score-vector sensitivity excludes the one PL fit per k that did not meet the extended stopping criterion; Borda is calculated on the same matched rows. The $k = 10$ result is not the primary endpoint test in Table A1.

Table A15. Borda tail-score convention sensitivity. Means across 40 seed–model rows using retained ranking payloads; the $k = 5$ rows are nested truncations, not a new run. No convention was selected by reference performance.

k	Tail convention	Assigned zero-based rank	Mean Kendall τ	Mean NDCG@10
3	Last disclosed rank	2.0	0.39220	0.66732
3	Next unseen rank (default)	3.0	0.40410	0.67874
3	Mean possible tail rank	15.0	0.40344	0.67847
3	Bottom rank	27.0	0.40397	0.67612
5	Last disclosed rank	4.0	0.42447	0.69551
5	Next unseen rank (default)	5.0	0.44497	0.70361
5	Mean possible tail rank	16.0	0.44841	0.69717
5	Bottom rank	27.0	0.44841	0.69306
10	Last disclosed rank	9.0	0.51508	0.68755
10	Next unseen rank (default)	10.0	0.52169	0.68236
10	Mean possible tail rank	18.5	0.51349	0.66104
10	Bottom rank	27.0	0.51349	0.64879

Table A16. Convergence-gated strict-Kendall sensitivity from the independent payload reconstruction. Each method is evaluated on the same retained seed/model rows whose extended PL-MM fit meets the 5,000-iteration/ 10^{-8} criterion. The $k = 5$ nested result is not repeated here. This is not the primary endpoint; client weights are reconstructed and one Borda discrepancy remains unresolved.

k	Valid rows	Mean Δ	95% t interval	Exact p	Positive	Negative/zero
3	39/40	+0.00013	[-0.00274, 0.00300]	0.984375	5	4/1
10	39/40	-0.00445	[-0.01863, 0.00973]	0.603516	5	5/0

References

- Wenjing Chen, Ruida Zhou, Chao Tian, and Cong Shen. On top-k selection from m-wise partial rankings via borda counting, 2022. URL <https://arxiv.org/abs/2204.05742>.
- Robert Detrano, Andras Janosi, W. Steinbrunn, Matthias Pfisterer, Jeffrey J. Schmid, S. Sandhu, Kenneth H. Guppy, Soo-Jin Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, 1989. doi: 10.1016/0002-9149(89)90524-9.
- Pietro Ducange, Francesco Marcelloni, Andrea Renda, and Fabrizio Ruffini. Federated learning of xai models in healthcare: A case study on parkinson’s disease. *Cognitive Computation*, 16:3051–3076, 2024. doi: 10.1007/s12559-024-10332-x.
- Ronald Fagin, Ravi Kumar, and Dandapani Sivakumar. Comparing top k lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003. doi: 10.1137/S0895480102412856.
- David R. Hunter. Mm algorithms for generalized bradley-terry models. *The Annals of Statistics*, 32(1):384–406, 2004. doi: 10.1214/aos/1079120141.
- Andras Janosi, W. Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart disease [dataset], 1989.
- Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurelien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2):1–210, 2021. doi: 10.1561/22000000083.
- Ashish Khetan and Sewoong Oh. Data-driven rank breaking for efficient rank aggregation. *Journal of Machine Learning Research*, 17(193):1–54, 2016. URL <https://jmlr.org/papers/v17/16-209.html>.
- Ashish Khetan and Sewoong Oh. Generalized rank-breaking: Computational and statistical tradeoffs. *Journal of Machine Learning Research*, 19(28):1–42, 2018. URL <https://jmlr.org/papers/v19/16-412.html>.
- R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.
- Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30, 2017.
- Jiaqi Ma, Xinyang Yi, Weijing Tang, Zhe Zhao, Lichan Hong, Ed Chi, and Qiaozhu Mei. Learning-to-rank with partitioned preference: Fast estimation for the plackett-luce model. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 928–936. PMLR, 2021. URL <https://proceedings.mlr.press/v130/ma21a.html>.
- 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*, volume 54, pages 1273–1282, 2017. URL <https://proceedings.mlr.press/v54/mcmahan17a.html>.

- Cristiano Migali, Gianmarco Genalti, Alberto Maria Metelli, and Marco Mussi. Learning to rank from incomplete rankings. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 88752–88788. PMLR, 2026. URL <https://proceedings.mlr.press/v306/migali26a.html>.
- A. C. Oksuz, A. Halimi, and Erman Ayday. Sok: Privacy attacks on machine learning via explainable ai, 2026. URL <https://arxiv.org/abs/2609.10627>.
- Yuangang Pan, Ivor W. Tsang, Weijie Chen, Gang Niu, and Masashi Sugiyama. Fast and robust rank aggregation against model misspecification. *Journal of Machine Learning Research*, 23(23):1–35, 2022. URL <https://jmlr.org/papers/v23/20-315.html>.
- Michael Pearce and Elena A. Erosheva. A unified statistical learning model for rankings and scores with application to grant panel review. *Journal of Machine Learning Research*, 23(210):1–33, 2022. URL <https://jmlr.org/papers/v23/21-1262.html>.
- Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975. doi: 10.2307/2346567.
- Nicola Rieke, James Hancox, Wenqi Li, Federica Milletari, Holger R. Roth, Shadi Albarqouni, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3:119, 2020. doi: 10.1038/s41746-020-00323-1.
- Syed Saifullah, David Mercier, Adriano Lucieri, Andreas Dengel, and Sheraz Ahmed. The privacy-explainability trade-off: unraveling the impacts of differential privacy and federated learning on attribution methods. *Frontiers in Artificial Intelligence*, 7:1236947, 2024. doi: 10.3389/frai.2024.1236947.
- Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 231–241, 2021. doi: 10.1145/3461702.3462533.
- Jack W. Smith, James E. Everhart, William C. Dickson, William C. Knowler, and Robert S. Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications and Medical Care*, pages 261–265. IEEE Computer Society Press, 1988.
- Hossein Azari Soufiani, David Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 360–368. PMLR, 2014. URL <https://proceedings.mlr.press/v32/soufiani14.html>.
- Heather L. Turner, Jeroen van Etten, David Firth, and Ioannis Kosmidis. Modelling rankings in r: the plackettluce package. *Computational Statistics*, 35:1027–1057, 2020. doi: 10.1007/s00180-020-00959-3.
- William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast Cancer Wisconsin (Diagnostic) [dataset]. UCI Machine Learning Repository, 1993. doi: 10.24432/C5DW2B. URL <https://archive.ics.uci.edu/dataset/17/breast-cancer>.

Li Zhao, Hui Xie, Li Zhong, and Yan Wang. Explainable federated learning scheme for secure healthcare data sharing. *Health Information Science and Systems*, 12:49, 2024. doi: 10.1007/s13755-024-00306-6.

Xi Zhao, Wei Zhang, Xinyuan Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–692, 2021.