# LOCO-FL: Scaling and Reference Choice for Bayesian Structural Analysis of Between-Center Heterogeneity in Federated Learning with Markov Boundaries

**András Millinghoffer**[1,2] MILLI@MIT.BME.HU
**Júlia Kisida**[1,2] KISIDA.JULIA.LILI@EDU.BME.HU
**Andrea Valek**[3,4] VALEK.ANDREA@SEMMELWEIS.HU
**Miklós Szabó**[3] SZABO.MIKLOS@SEMMELWEIS.HU
**Péter Antal**[1,2,*] ANTAL@MIT.BME.HU

[1] *Budapest University of Technology and Economics, Budapest, Hungary*
[2] *E-Group ICT Software Zrt., Budapest, Hungary*
[3] *Semmelweis University, Budapest, Hungary*
[4] *Heim Pál Children's Hospital, Budapest, Hungary*
[*] *Corresponding author.*

**Editors:** Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

## Abstract

Multicenter quality-improvement programs still struggle to explain between-center heterogeneity using outcome measures alone. We investigate a Bayesian network method that focuses on Markov boundaries to characterize this heterogeneity at the structural level of risk models with multiple outcomes. Working with the full structure posterior rather than a point estimate allows the study of a phenomenon that point-identification learners overlook: under proper regularization, the resolvable Markov-boundary complexity grows with sample size. A naive comparison across centers, therefore, mixes sample-size induced model properties with genuine structural differences. We therefore study a global reference versus leave-one-center-out (LOCO) references together with sample-size correction, and show that both can be analyzed within a single reweighting of the per-family BD sufficient statistics, achieved exactly by adding the $\beta$-scaled LOCO counts as pseudocounts to the score. We investigate augmenting small centers to a larger, common effective size using the LOCO sample. We present two exact, privacy-preserving (PP) Bayesian federated learning schemes: a single global-reference scheme and a LOCO-reference scheme (LOCO-FL). We show that the LOCO-FL scheme supports per-center reference comparison and per-center scaling under the same institutional-privacy guarantees. Because the LOCO aggregate excludes a center's own data, the same LOCO counts can also serve as an informative prior for that center's local inference, without self-inclusion. We choose the scale of this prior empirically, as the value that most reduces the posterior spread, which reveals a trade-off between variance reduction and loss of the between-center signal. On synthetic data with known between-center differences, the method recovers these differences above the within-posterior spread. We report results for an eight-center Hungarian very preterm cohort (2,939 patients), measuring heterogeneity by posterior Jaccard distances between LOCO and global references. Results confirm the need to scale to a common effective sample size, the utility of a single global reference for coarse comparison, and that an empirically scaled LOCO prior reduces posterior spread.

**Keywords:** Bayesian network structure learning, multi-target Markov boundaries, privacy-preserving federated learning, leave-one-center-out references, power priors, effective sample size, multicenter heterogeneity

© 2026 A. Millinghoffer, J. Kisida, A. Valek, M. Szabó & P. Antal.

## 1. Introduction

Federated learning (FL) integrates distributed healthcare data without centralizing patient records (Sharma and Guleria, 2023) and is the natural setting for multicenter structural analysis. Two challenges remain. The first is institutional: between-center heterogeneity is the main concern of quality-improvement (QI) programs, which aim to give useful feedback to the participating centers while protecting their privacy, that is, without revealing their outcome measures or their raw data. The second challenge is methodological: the predictive methods that dominate FL, even with recent advances in explainable AI (Schwalbe and Finzel, 2024; Albahri et al., 2023), cannot capture the multivariate relevance patterns behind heterogeneity. They can only re-rank centers using aggregate metrics, leaving the heterogeneity itself unexplained (van der Linde et al., 2024).

We address this gap with a Bayesian PGM-based approach: rather than re-ranking centers based on outcome metrics, we analyze which multivariate relevance patterns differ across centers and which predictors drive the differences. Probabilistic graphical models (PGMs) and the concept of Markov boundary provide a principled framework for multivariate relevance (Pearl, 1988; Cooper et al., 1997; Koller and Sahami, 1996; Friedman and Koller, 2000; Tsamardinos and Aliferis, 2003; Theng and Bhoyar, 2024), which extends naturally to the multi-target setting (Antal et al., 2008; Marx et al., 2016; Liu and Liu, 2018; Wu et al., 2023). Moreover, Bayesian inference for PGMs yields posteriors over model structures rather than point estimates, enabling full quantification of uncertainty in the multivariate relevance relations and their context (Friedman and Koller, 2003; Antal et al., 2008). The present work develops *exact* Bayesian FL by sampling over graph structures: between-center heterogeneity is preserved in the posterior representation rather than collapsed onto a single model, and ordering-MCMC provides asymptotically exact exploration of the structure posterior rather than the point estimate returned by identification-style learners.

In this paper, we address the following questions:

- Q1 How can the uncertainty in structural relevance be measured and compared across centers when multiple outcomes are involved?
- Q2 How can sample-size effects be adjusted for, so that between-center comparison reflects genuine heterogeneity across populations of different sizes?
- Q3 How can privacy-preserving (PP) federated learning compute LOCO references for local quality control operations, and how does it accommodate the per-center scaling required by Q2?
- Q4 Can the LOCO statistics serve as an informative prior on a center's local inference, and how can this be implemented in PP-FL?

Together, these questions motivate a multivariate, uncertainty-aware characterization of between-center variation at the structural level that complements outcome-level benchmarking, and is particularly relevant for the care of preterm newborns in NICUs, where multiple correlated outcomes, rare adverse events, small per-center samples, and well-documented between-center variation in mortality and morbidity (Edwards et al., 2019; Nador et al., 2025) all coexist. Fig. 1 summarizes the methodology.

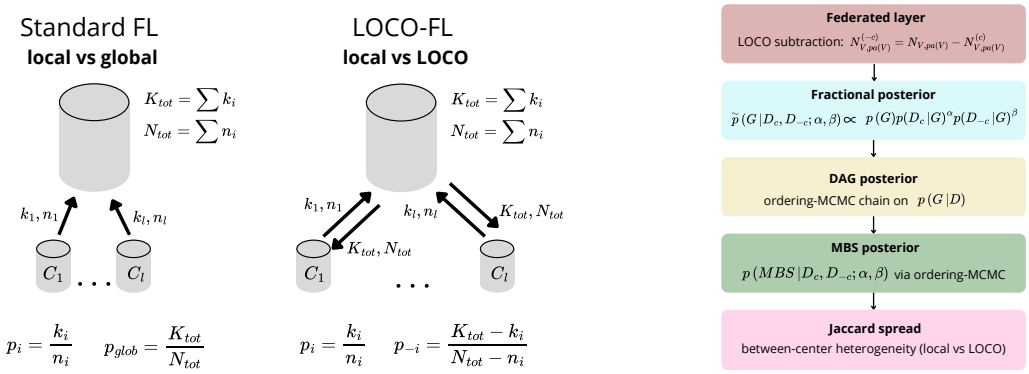



Figure 1: **(Left)** Standard FL compares each center's local rate $p_i$ to the global rate $p_{glob}$, which already includes that center. LOCO-FL broadcasts the aggregate, so each center reconstructs a leave-one-out reference $p_{-i}$ via self-subtraction, thereby removing the self-inclusion bias. **(Right)** The per-center LOCO-FL pipeline runs entirely locally: the leave-one-out statistics feed a fractional posterior that weights the center's own data by $\alpha$ and the leave-one-out data by $\beta$; ordering-MCMC then gives the Markov boundary posterior, and finally the between-center Jaccard spread.

## 2. Background and Related Work

The proposed method combines Bayesian structure learning over Markov boundaries, fractional posteriors and power priors, and exact federated inference through sufficient-statistic sharing, which we review before relating LOCO-FL to prior work on cross-silo personalization, leave-one-site-out evaluation, and mixture-aware structure learning. *Bayesian network structure learning and Markov boundaries:* Bayesian inference over directed acyclic graphs (DAGs) provides a principled framework for multivariate relevance analysis through the Markov boundary (Pearl, 1988; Koller and Sahami, 1996; Friedman and Koller, 2000; Tsamardinos and Aliferis, 2003), with extensions to the multi-target setting (Antal et al., 2008; Marx et al., 2016; Wu et al., 2023). Under decomposable scores, the score of a DAG factorizes over per-family sufficient statistics. This makes both exact federated implementation and per-family power-prior reweighting possible in principle, independently of score equivalence. We use the Bayesian Dirichlet (BD) family (Cooper and Herskovits, 1992; Heckerman et al., 1995), which includes the Cooper–Herskovits/K2 score, the uniform-pseudocount variant used here, and the score-equivalent BDeu specialization. Markov-boundary-set (MBS) posteriors can be estimated efficiently by ordering-MCMC (Friedman and Koller, 2003; Antal et al., 2008). *Fractional posteriors and power priors:* The fractional posterior $\tilde{p}(\theta \mid D; \alpha) \propto p(\theta)\, p(D \mid \theta)^{\alpha}$, $\alpha \in [0, 1]$, has a long history in Bayesian robustness (Walker and Hjort, 2001) and as the power-prior construction of Ibrahim and Chen (2000), with recent theoretical work establishing its concentration and merging properties (Bhattacharya et al., 2019). The two-component variant $\tilde{p}(\theta \mid D_1, D_2; \alpha, \beta) \propto p(\theta)\, p(D_1 \mid \theta)^{\alpha}\, p(D_2 \mid \theta)^{\beta}$ is standard in the historical-borrowing literature (Neuenschwander et al., 2009; Hobbs et al., 2011), where $D_2$ is a historical control and $\beta$ controls borrowing

strength. *Federated Bayesian inference and sufficient-statistic sharing:* The starting point for exact Bayesian federated structure learning is Ma and Sivakumar (2006), who observe that decomposable scores require only the per-family sufficient statistics, so a single secure-aggregation round suffices for exact federated structure scoring. More recent work has extended this idea to the parameter-posterior setting: Jonker et al. (2024, 2025) develop *Bayesian Federated Inference* (BFI), in which each center runs a local Bayesian analysis and a single aggregation round at the server returns the pooled posterior to all clients. One-shot Bayesian FL more broadly (Neiswanger et al., 2013; Al-Shedivat et al., 2020; Jhunjhunwala et al., 2024; Liu et al., 2024; Talpini et al., 2025) uses the same idea: clients send approximate local posteriors or Fisher information once, and the server aggregates into a global posterior. *Cross-silo personalization and the self-inclusion problem:* A separate line of cross-silo FL aims not at a global model but at per-silo personalization. Liu et al. (2022) propose mean-regularized multi-task learning (MR-MTL) under silo-specific differential privacy, in which each silo trains its own model with a regularization term toward the *mean of all silo models*. The construction documents the value of per-silo aggregate-as-reference comparisons, but the reference is the global mean and therefore includes the silo's own contribution. Subtracting a silo's own contribution from a returned global aggregate also appears, in a different role, in iterative-optimization FL, where SCAFFOLD-style control variates (Karimireddy et al., 2020) use it to reduce the variance of stochastic gradients. *Leave-one-site-out in multicenter FL:* Leave-one-site-out (LOSO) is well-established as an evaluation protocol in multicenter healthcare ML (Kazlouski et al., 2025): a model is trained on $C-1$ sites and tested on the held-out site, reporting cross-site generalization. As an *evaluation* protocol, it does not return a per-site reference for downstream analysis, and a naive implementation requires $C$ separate aggregation rounds (one per held-out site). *Mixture-aware BN learning and personalized federated structure:* Pooling data from heterogeneous centers under a single BN learner is a known identifiability hazard: the recovered graph converges to a pseudo-observational equivalence class that conflates between-center heterogeneity with sample-size variation. Mixture-aware BN learning approaches recover subpopulation-specific structures (Shaddox et al., 2020) but typically require centralized access to labeled data and scale poorly in the number of centers. The recent NOTEARS-PFL framework (Liu et al., 2025) learns per-site Bayesian networks under a sparse-group-lasso penalty within an ADMM-style federated optimization.

## 3. Methods

We describe the cohort, the structural summaries, the reweighting, and the federated scheme that computes it.

### 3.1. Data

The data come from the Hungarian national perinatal registry (Országos Kórházi Főigazgatóság (OKFŐ), 2023), restricted to the 2014–2016 period during which an expanded, EuroNeonNet-aligned protocol (Haumont et al., 2020) was in operation at the participating institutes. Only newborns born before the 33rd gestational week were retained, on grounds of clinical relevance. Variables were screened for excessive missingness, leaving 51 variables with

average and maximum per-variable missingness rates of 2.04% and 31%, respectively. Missingness was handled by univariate imputation from each variable's marginal discrete distribution, which preserves the multivariate dependency structure needed for structure learning. Nine variables were originally continuous and were discretized. Following Nador et al. (2025), ten outcome variables were selected for joint multi-target analysis. Institutes were filtered at $n_c > 200$, a sensible lower limit for Bayesian structural analysis over 51 discrete variables, below which the structure posterior remains largely prior-dominated (cf. Fig. 3). The eight remaining Level-III NICUs comprise 2,939 newborns, yet per-center sizes still span a nearly fourfold range (up to $n_c = 742$), which alone can confound per-center comparison (Q2).

### 3.2. LOCO-FL with MBS Posteriors and Sample-Size Reweighting

Let $C$ denote the number of centers. We write $D_c$ for the data of center $c \in \{1, \ldots, C\}$ and $D_{-c} = \bigcup_{c' \neq c} D_{c'}$ for the leave-one-center-out (LOCO) sample. Per-center sample sizes are $n_c = |D_c|$ and $N_{-c} = |D_{-c}|$; the pooled sample size is $N = \sum_c n_c$. We work with discrete Bayesian networks scored by the Bayesian Dirichlet (BD) score (Cooper and Herskovits, 1992; Heckerman et al., 1995) under a uniform Cooper–Herskovits (K2-style) prior with a fixed pseudocount $\eta$ per cell ($\alpha^0_{V,j,k} = \eta$ for all $(V, j, k)$; we use $\eta = 2$), sampled by ordering-MCMC (Friedman and Koller, 2003). Unlike its score-equivalent BDeu specialization, which sets $\alpha^0_{V,j,k} = \alpha_{\mathrm{ESS}}/(r_V q_V)$, the uniform-pseudocount BD score is not score-equivalent. We adopt it deliberately, because the per-cell BDeu prior vanishes as parent sets grow, making BDeu unstable and strongly $\alpha_{\mathrm{ESS}}$-sensitive for the rare adverse events and small per-center samples studied here (Silander et al., 2018), whereas a fixed per-cell pseudocount avoids this problem. The lack of score equivalence does not affect the structural summaries below: the Markov boundary set (MBS) is invariant across an observational equivalence class, so its push-forward is the same regardless of which equivalence-class member the score happens to prefer. Under a continuous prior over the network parameters, faithfulness holds with probability one, so each per-target $\mathrm{MB}(Y_j; G)$ is the unique Markov boundary — the minimal Markov blanket — not merely a blanket. For a target set $\mathbf{Y} = \{Y_1, \ldots, Y_q\}$ of $q$ outcomes, the Markov-boundary set (MBS) of $\mathbf{Y}$ in a DAG $G$ is the union of these per-target boundaries, $\mathrm{MBS}(\mathbf{Y}; G) = \bigcup_{j=1}^{q} \mathrm{MB}(Y_j; G)$, and its complexity is the cardinality $|\mathrm{MBS}(\mathbf{Y}; G)|$. The posterior MBS distribution at a given reference is the push-forward of the structure posterior through $G \mapsto \mathrm{MBS}(\mathbf{Y}; G)$.

For each reference posterior $\pi$ over DAGs (per-center, LOCO, global, or any reweighted variant), the structural posterior is summarized at the level of the MBS rather than at the full-graph. This compresses the representation to the multi-target relevance information of interest. As a single scalar of posterior spread we use the *expected pairwise Jaccard distance* on the MBS, $\bar{J}(\pi) = \mathbb{E}_{S,S' \sim \pi}\, d_J(S, S')$, where $d_J(S, S') = |S \triangle S'|/|S \cup S'|$ and $S, S'$ are independent posterior draws; alongside it we report the unnormalized symmetric difference $|S \triangle S'|$. The same construction compares two reference posteriors, $\bar{J}(\pi_a, \pi_b) = \mathbb{E}_{S \sim \pi_a,\, S' \sim \pi_b}\, d_J(S, S')$, which we call the between-posterior distance. Note that $\bar{J}(\cdot, \cdot)$ is not itself a metric between posteriors: $\bar{J}(\pi, \pi)$ equals the within-posterior spread rather than zero, which is why between-posterior values are always interpreted relative to the within-posterior spreads.

To separate sample-size confounding from genuine structural heterogeneity, we weight the local data $D_c$ and the LOCO data $D_{-c}$ by independent weights $(\alpha, \beta)$. Conceptually, this is the fractional structure posterior

$$\tilde{p}(G \mid D_c, D_{-c};\, \alpha, \beta) \;\propto\; p(G)\, p(D_c \mid G)^{\alpha}\, p(D_{-c} \mid G)^{\beta}, \qquad (\alpha, \beta) \in [0,1]^2, \tag{1}$$

but under the BD score the marginal likelihood of a DAG factorizes over per-family Dirichlet–multinomial terms, each a function only of the family count tensor. We therefore reweight *these count tensors* and re-score *exactly*. For each family $(V, \mathrm{pa}_V)$ define the effective count tensor

$$\tilde{N}^{(c)}_{V,\mathrm{pa}_V}(\alpha, \beta) \;=\; \alpha\, N^{(c)}_{V,\mathrm{pa}_V} \;+\; \beta\, N^{(-c)}_{V,\mathrm{pa}_V}, \tag{2}$$

and score $G$ by the exact BD marginal likelihood at $\tilde{N}^{(c)}$ with the uniform BD prior $\alpha^0_{V,j,k} = \eta$ ($\eta = 2$; $r_V$ states of $V$, $q_V$ parent configurations, $\alpha^0_{V,j} = \sum_k \alpha^0_{V,j,k} = r_V \eta$, $\tilde{N}_{V,j} = \sum_k \tilde{N}_{V,j,k}$). The induced parameter posterior is the exact Dirichlet $\mathrm{Dir}\big(\alpha^0 + \alpha N^{(c)} + \beta N^{(-c)}\big)$, in which the $\beta$-scaled LOCO counts act as prior pseudocounts that augment the uniform BD prior. Equation (1) recovers the local posterior at $(\alpha, \beta) = (1, 0)$, the LOCO posterior at $(0, 1)$, and the pooled posterior at $(1, 1)$, and is computable from the count tensors alone, hence federated. Sample-size matching targets a common comparison level $N_0$: a center is down-weighted only when it exceeds $N_0$, and a center below $N_0$ retains all its observations and is augmented from its LOCO sample:

$$\alpha_c = \min\Big(1,\, \frac{N_0}{n_c}\Big), \qquad \beta_c = \frac{\max\big(0,\, N_0 - n_c\big)}{N_{-c}}, \tag{3}$$

under which the effective size $\alpha_c n_c + \beta_c N_{-c} = N_0$ for every center.

For each variable $V$ and parent set $\mathrm{pa}_V$ the BD score depends on $D$ only through the count tensor $N^{(c)}_{V,\mathrm{pa}_V}$ of $(V, \mathrm{pa}_V)$ joint configurations in center $c$. We cap the in-degree at $k = 4$, a sensible limit given the available data. This reduces the number of candidate parent sets per node from $2^{n-1}$ to $\sum_{j=0}^{k} \binom{n-1}{j} = \Theta(n^k)$; however, due to marginalization and parent-child equivalence, only the count tensors for the unique $(k+1)$-sized family sets need to be shared. The exchange remains exponential, a limit to scaling in $k$, however, since each tensor has up to $r^{k+1}$ cells for arity-$r$ variables: at $n = 51$ this amounts to approximately $2.35 \times 10^6$ families and $5.71 \times 10^8$ counters in the single aggregation round, assuming $r = 3$ for every variable (approximately 1.1 GB data to transfer using 2-byte integers). The pooled count tensor over all centers is $N_{V,\mathrm{pa}_V} = \sum_c N^{(c)}_{V,\mathrm{pa}_V}$, and per-family scores are additive in these counts. In standard privacy-preserving federated inference (Ma and Sivakumar, 2006; Jonker et al., 2024), each center sends its per-family count tensor once under a secure-aggregation protocol (e.g., Bonawitz et al., 2017); the server forms the pooled count $N_{V,\mathrm{pa}_V}$ and returns it to all centers, which then independently run identical MCMC on the same pooled posterior. This yields a single global posterior, identical at every center. The LOCO-FL scheme replaces the final step with a single arithmetic operation at each client. After receiving the pooled count $N_{V,\mathrm{pa}_V}$, each center $c$ computes its leave-one-center-out count by local subtraction, which uses only the global aggregate (already returned) and the center's own private count (which never left the center). Each center then runs an independent MCMC chain on $p(G \mid D_{-c})$, the LOCO posterior, using $N^{(-c)}_{V,\mathrm{pa}_V}$ as sufficient

statistics, in parallel with an MCMC chain on its local posterior $p(G \mid D_c)$. The per-center LOCO-vs-local comparison is then carried out entirely locally via the Jaccard distance, with no further communication. The reweighting is applied at this stage by forming the effective counts $\tilde{N}^{(c)}_{V,\mathrm{pa}_V}(\alpha, \beta)$ of Eq. (2) before scoring. The pseudocode is given in Algorithm 1.

The LOCO sufficient statistics produced by Algorithm 1 can be re-used in a second way: instead of serving only as a reference against which to contrast the local posterior, they can act as an *informative prior* on the local inference itself. In the reweighting this corresponds to taking $\beta > 0$ with $\alpha = 1$, where we investigate the effect of $\beta$ on detection of between-center differences.

To validate the methodology on data with known generating structures, 5 independent synthetic datasets were sampled for each center matching the center's original sample size from its maximum a posteriori (MAP) model with parameters estimated from that center's data. On each synthetic dataset we ran 5 independent MCMC chains with an effective data size of 200, so each synthetic posterior averages $5 \times 5$ samplings. The resulting synthetic MBS posteriors were compared with both the generating model structures and the corresponding real-data posteriors.

## 4. Results

We report results for varying sample sizes on the eight-center cohort under the following configurations: the pooled (global) posterior on all data; the eight leave-one-center-out posteriors $p(G \mid D_{-c})$; the eight local posteriors $p(G \mid D_c)$; and their LOCO combined variants. We report the within-posterior spread $\bar{J}(\pi)$ and the between-posterior distance $\bar{J}(\pi_a, \pi_b)$ on the MBS, under both the normalized Jaccard metric and its unnormalized symmetric-difference (set-cardinality) counterpart. We assessed convergence of the MCMC samples using the Geweke and Gelman–Rubin tests for the posteriors of all possible pairwise Markov-boundary membership (MBM) relations. On real data, with an effective sample size of 200, 75.15% of all the Geweke tests conducted showed within-chain convergence (i.e. the corresponding p-value was greater than 0.05), whereas 95.17% of all the Gelman–Rubin statistics were less than 1.01, showing sufficient inter-chain concordance. On synthetic data our method recovers the known structural differences: the between-posterior distances estimated from the synthetic data follow the true structural distances between the generating networks (Spearman $\rho = 0.66$, $p = 1.3 \cdot 10^{-4}$), and also agree with the between-posterior distances estimated from the real data ($\rho = 0.72$, $p = 10^{-5}$, see Fig. 2 and Fig. 4). The differences between the generating networks are nevertheless larger than those between the corresponding MBS posteriors, and they agree with the real-data posteriors the least of the three ($\rho = 0.48$, $p = 0.01$). With convergence and the recovery of known differences thus established, the remaining results address the four questions of Section 1 in turn.

**Structural uncertainty over sample sizes (Q1)** The level of structural uncertainty and its dependence on sample size is shown in Fig. 3 left. To illustrate the underlying change in model complexity, Fig. 3 shows the expected MBS cardinality (top middle) and the number of free parameters of the Markov-boundary subgraph (MBG) spanned by the MB set (bottom middle). Although the uncertainty is high and depends strongly on sample size, Fig. 3 right shows the distances to the aggregated global and LOCO references.

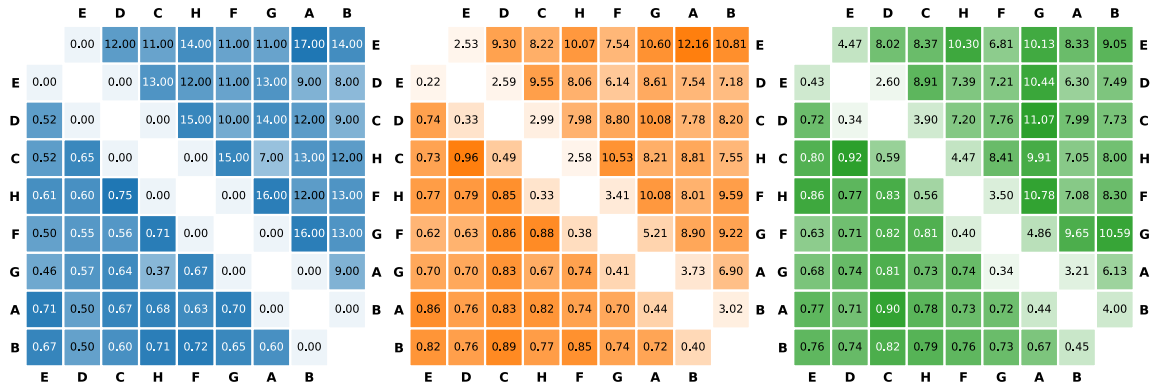



Figure 2: Pairwise distances between centers at a matched effective sample size of 200. In each panel, the lower triangle shows the Jaccard distance and the upper triangle the symmetric difference. The diagonal of each triangle gives the within-posterior spread in that triangle's metric (zero in the left panel, where each center is a single MAP structure). *Left:* between the centers' maximum a posteriori (MAP) models. *Middle:* between MBS posteriors learned on synthetic datasets generated from those MAP models. *Right:* between the centers' MBS posteriors learned on their real data.

**Sample-size matched comparison (Q2)** Owing to this sample-size dependence, between-center heterogeneity must be compared at a common sample size, which is limited by the smaller centers. Our data include only eight centers with the largest sample sizes after filtering. Fig. 5 shows the between-posterior distances to the global and to the leave-one-center-out references, at the matched (200) and full sample sizes.

**Global vs. LOCO references (Q3)** Besides sample size confounding, the global reference posterior produced by a standard FL is limited by self-inclusion for per-center prior incorporation and comparison. The proposed LOCO-FL scheme provides efficiently computable LOCO posteriors, allowing more precise comparison; see Fig. 5.

**LOCO prior as an informative prior (Q4)** Given the sample-size dependence established in Q1 and Q2, and the availability of LOCO posteriors from LOCO-FL, we re-use the LOCO posterior as an informative prior on a center's local inference, taking the largest center as the worked example. Varying both the local base size (200/300/400/orig) and the LOCO-prior scale, we track the posterior spread and the distance to the pooled reference and the center's original MBS posterior (Figs. 6–7).

## 5. Discussion

We discuss the findings in relation to the four questions in Section 1 and conclude with limitations.

*Significant and growing structural uncertainty (Q1):* The learning curves show that the residual uncertainty is substantial ($\bar{J} \approx 0.3$ even at the pooled scale): the expected pairwise

Jaccard spread $\bar{J}(\pi)$ falls with sample size without plateauing (Fig. 3, top left), while in the unnormalized symmetric-difference metric the spread grows (bottom left), because the resolvable model complexity itself grows with sample size (MB-set size and MBG parameter count; middle panels). Two centers with different sample sizes therefore sit at different points on these curves, so part of any raw structural difference reflects the difference in sample size, through the implicit complexity regularization, rather than a genuine between-center difference.

*Exact sample-size matching (Q2):* The exact reweighting corrects this by matching effective sample size: at a matched size of 200 the between-center Jaccard distances ($\approx 0.63$–$0.92$) clearly exceed the within-center spreads ($\approx 0.34$–$0.59$; Fig. 2, right panel, off-diagonal vs. diagonal entries of the lower triangle), so the residual heterogeneity is credible above the per-center noise floor. Own-data reweighting can only equalize centers *downward*; thus, the largest reachable size is the minimum per-center size. The remedy is the asymmetric rule of Eq. (3): trim large centers and augment small ones from the LOCO sample, so that comparison runs at a common, possibly larger, effective size; this is examined in relation to Q4.

*Federated LOCO posteriors (Q3):* Fig. 5 shows that the global reference gives a reasonably good ranking of center deviations, while the LOCO references, which avoid self-

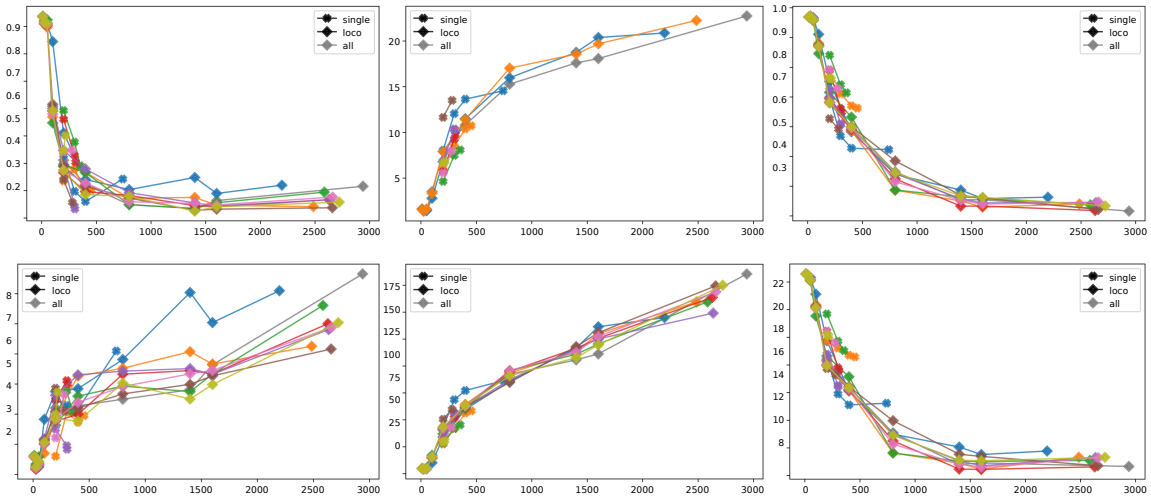



Figure 3: Across all panels, the MBS posterior is characterized as a function of (sub)sample size for the pooled (*all*), per-center (*single*) and LOCO (*loco*) posteriors, in terms of its within-posterior spread (*top left*: Jaccard distance, *bottom left*: symmetric difference), its expected complexity (*top middle*: MBS cardinality, *bottom middle*: free-parameter dimension of the MBG), and its deviation from the pooled reference posterior (*top right*: Jaccard distance, *bottom right*: symmetric difference). Gray diamonds denote values corresponding to the pooled dataset, whereas diamonds and crosses in various colors denote points corresponding to the LOCO and per-center posteriors of the same institute, respectively.

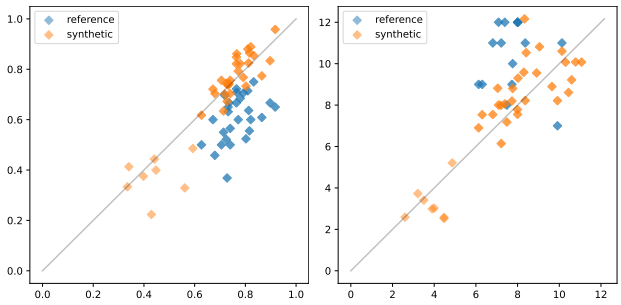



Figure 4: Comparison of pairwise distances between centers in different scenarios: the x-axis shows the between-posterior distances among the centers' real-data posteriors at matched effective size 200; the y-axis shows the expectation of the structural distances between the reference model structures (blue) and posterior samples from the synthetic data generated from the reference models (orange). The left panel shows Jaccard distances, whereas the right panel shows the symmetric differences.

| | F | G | E | B | D | A | H | C |
|---|---|---|---|---|---|---|---|---|
| all, 200 | 0.585 | 0.652 | 0.506 | 0.733 | 0.730 | 0.752 | 0.787 | 0.708 |
| all, full | 0.607 | 0.592 | 0.523 | 0.760 | 0.663 | 0.729 | 0.649 | 0.714 |
| loco, full | 0.561 | 0.619 | 0.506 | 0.777 | 0.660 | 0.722 | 0.633 | 0.722 |
| loco, 200 | 0.622 | 0.658 | 0.722 | 0.745 | 0.763 | 0.766 | 0.790 | 0.845 |

| | F | A | E | C | D | B | H | G |
|---|---|---|---|---|---|---|---|---|
| all, 200 | 6.134 | 8.048 | 5.490 | 6.806 | 8.062 | 8.427 | 8.732 | 9.485 |
| all, full | 14.431 | 17.578 | 13.233 | 17.039 | 16.581 | 18.094 | 15.610 | 15.288 |
| loco, full | 12.457 | 16.332 | 12.020 | 16.685 | 16.202 | 19.094 | 14.822 | 16.036 |
| loco, 200 | 6.661 | 8.361 | 8.364 | 8.448 | 8.613 | 8.788 | 8.938 | 9.741 |

Figure 5: Distance of each center's MBS posterior to the global (*all*) and leave-one-center-out (*loco*) references, at the matched (200) and full sample sizes; Jaccard distance (*left*) and symmetric difference (*right*).

inclusion, give a sharper range of deviations, illustrating that the LOCO-FL method efficiently provides LOCO posteriors for the centers, both for comparison and prior incorporation. The privacy and security advantages of LOCO-FL are as follows: First, with secure aggregation, the server observes only the global sum $N_{V,\mathrm{pa}_V}$, never an individual count $N^{(c)}_{V,\mathrm{pa}_V}$. Second, it requires a single secure-aggregation round and one broadcast, rather than the $C$ rounds a naive per-held-out-center construction would require. Additionally, note that a server with the global aggregate and a LOCO reference could recover that center's statistics. Client-side subtraction is the only construction in which the single (global aggregate, own count) pair that is ever co-located sits at the center that already owns its count. The $(\alpha, \beta)$ reweighting of Q2 inherits this locality and is applied entirely client-side.

*Effect of LOCO priors (Q4):* The augmented posterior converges to the pooled reference as the weight for the incorporated LOCO prior grows (Fig. 6). Furthermore, for most base sample sizes, the posterior spread first falls and then rises (Fig. 7). Thus, borrowing from $D_{-c}$ biases each posterior toward the pooled structure, tightening the per-center spread while attenuating the between-center signal. This can be expected in the small-sample regime, where the raw spread exceeds the between-center distance. This also connects

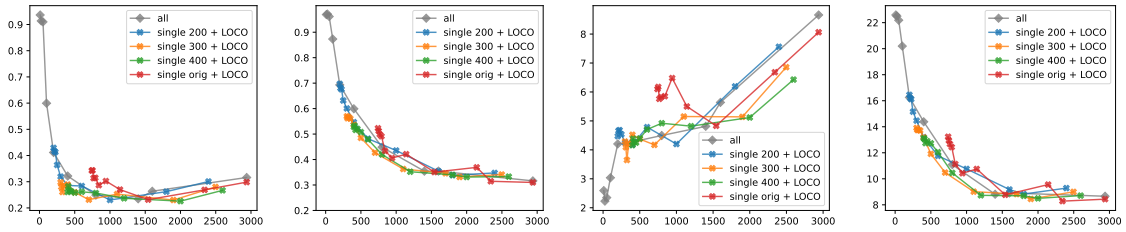



Figure 6: Local-plus-LOCO augmentation for the largest center, for several local base sizes (200/300/400/orig), each augmented by a LOCO prior of increasing scale ($x$-axis: total effective size). From left to right, the first and second panels show the within-posterior spread and the between-posterior distance from the pooled reference, quantified by Jaccard distance. The third and fourth panels show the corresponding measures using symmetric difference.

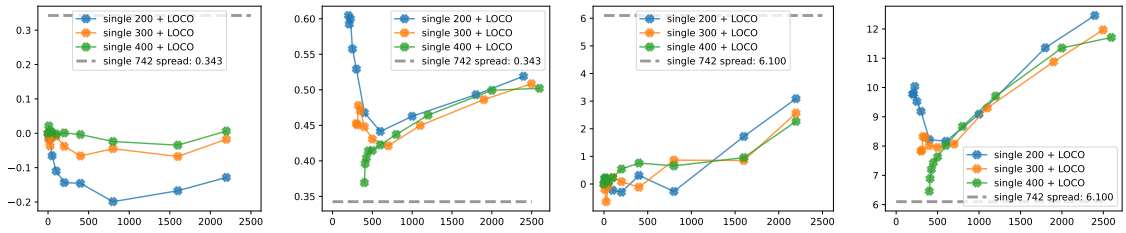



Figure 7: Effect of the scaled LOCO prior on the original local posterior for the largest center. The first and third panels show the change in spread, quantified by Jaccard distance and symmetric difference, respectively, between posteriors from the augmented datasets and the original local dataset. The $x$-axis indicates the difference in data size, i.e., the effective number of LOCO samples added. The second and fourth panels show the between-posterior distances between the augmented datasets and the original local posterior, with the $x$-axis indicating the combined data size. Colors denote different effective local base sizes; the dashed gray line indicates the spread of the posterior for the original local dataset ($n_c = 742$).

to the "larger effective size" objective of Q2: since resolvable complexity grows with $n$, comparing at a common larger size via LOCO augmentation resolves finer structure than the minimum-size floor permits. Note that the LOCO evidence enters as exact Dirichlet pseudocounts rather than a power on the integrated posterior, so the construction is exact under decomposable scores, avoiding the normalizing-constant and boundary-degeneracy issues that complicate selection of the borrowing strength in structure-level power priors.

*Limitations:* The present work is limited by the following four factors. First, the LOCO-prior scaling reported here is screened *empirically* for the value that most reduces the posterior spread (Fig. 7), not selected by a principled criterion. Minimizing within-center spread and maximizing between-center separation coincide only approximately, since bor-

rowing that tightens a center's posterior also pulls it toward the pool and attenuates the very between-center signal to be detected. A principled per-center choice of the borrowing strength, whether by a marginal-likelihood (empirical-Bayes) criterion or by a detection-targeted criterion that maximizes the between-center distance relative to the within-center spread, is the natural next step. Second, the screen is demonstrated on a single worked center (the largest), at a fixed local effective size. A systematic per-center profile across all eight centers, together with how the spread-minimizing scaling varies with each center's sample size and case-mix divergence from the LOCO pool, remains to be established, and the regime of practical interest is a *bounded* one, since unbounded borrowing converges every center to the pooled reference and dissolves the heterogeneity (Figs. 6–7). Third, the analysis rests on a single national cohort; the complexity-growth and augmentation findings require replication on an independent registry and a sensitivity analysis over the prior pseudocounts. Fourth, the cost of the single-round exactness of LOCO-FL is the high communication volume (Section 3.2): scaling beyond the present $n$ and in-degree limit requires candidate-parent screening and the compression of the count tensors by context-specific independences, e.g., decision-tree local structures (Boutilier et al., 1996; Friedman and Goldszmidt, 1996; Chickering et al., 1997).

## 6. Conclusion

To understand and improve the quality of care in multicenter programs, we examine risk factors using a Bayesian, systems-based approach, studying the full posterior over Markov-boundary sets. This uncertainty-preserving analysis shows that the statistically resolvable model complexity does not saturate within the observed sample-size range, so a naive comparison of center-specific risk models and their posteriors mixes sample-size induced model properties with genuine between-center differences. A single two-parameter reweighting of the per-family BD sufficient statistics independently scales a center's own counts and incorporates the leave-one-center-out counts as pseudocounts; it contains the local, LOCO, and pooled posteriors as special cases and provides an exact sample-size correction for Bayesian network structure inference, computable from the shared count tensors alone. We showed that matching the centers to a common, realistically attainable effective sample size raises the between-center distances above the within-center posterior spreads. We also showed that the global reference suffices for a coarse ranking of between-center differences, whereas LOCO references support fine per-center comparison. Additionally, LOCO references can serve as informative priors, reducing a center's posterior spread, although only bounded borrowing preserves the between-center signal. Finally, the proposed LOCO-FL scheme obtains the LOCO references and the per-center scaling from a single secure-aggregation round via client-side subtraction, under the same institutional-privacy guarantee as standard federated scoring, and with a by-design safeguard: the global aggregate and a LOCO reference are never co-located at the server, nor at any party other than the center itself.

## Acknowledgments

This research was funded by the National Research, Development and Innovation Fund of Hungary under grant OTKA K139330. The authors thank the participating Hungarian

Level-III NICUs and the Hungarian Society of Perinatology and Obstetric Anesthesiology for access to the registry data. The study was approved by the Ethics Committee of the National Medical Research Council (ETT-TUKEB BM/8893-1/2023).

## References

Maruan Al-Shedivat, Jennifer Gillenwater, Eric Xing, and Afshin Rostamizadeh. Federated learning via posterior averaging: A new perspective and practical algorithms. *arXiv preprint arXiv:2010.05273*, 2020.

Ahmed Shihab Albahri, Ali M Duhaim, Mohammed A Fadhel, Alhamzah Alnoor, Noor S Baqer, Laith Alzubaidi, Osamah Shihab Albahri, Abdullah Hussein Alamoodi, Jinshuai Bai, Asma Salhi, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information fusion*, 96:156–191, 2023.

Péter Antal, András Millinghoffer, Gábor Hullám, Csaba Szalai, and András Falus. A Bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. In *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008*, volume 4 of *Proceedings of Machine Learning Research*, pages 74–89. PMLR, 2008.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1175–1191, 2017. doi: 10.1145/3133956.3133982.

Craig Boutilier, Nir Friedman, Moisés Goldszmidt, and Daphne Koller. Context-specific independence in Bayesian networks. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 115–123. Morgan Kaufmann, 1996.

David Maxwell Chickering, David Heckerman, and Christopher Meek. A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 80–89. Morgan Kaufmann, 1997.

G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

G. F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9:107–138, 1997.

Erika M Edwards, Danielle EY Ehret, Roger F Soll, and Jeffrey D Horbar. Vermont Oxford Network: a worldwide learning community. *Translational Pediatrics*, 8(3):182–192, 2019.

N. Friedman and D. Koller. Being Bayesian about network structure. In *Proc. of the 16th UAI Conf.*, pages 201–211, 2000.

N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.

Nir Friedman and Moisés Goldszmidt. Learning Bayesian networks with local structure. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 252–262. Morgan Kaufmann, 1996.

Dominique Haumont, Neena Modi, Ola D Saugstad, Rosine Antetere, Cuong NguyenBa, Mark Turner, Kate Costeloe, and Willem Aelvoet. Evaluating preterm care across Europe using the eNewborn European network database. *Pediatric research*, 88(3):484–495, 2020.

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

Brian P. Hobbs, Bradley P. Carlin, Sumithra J. Mandrekar, and Daniel J. Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011.

Joseph G. Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Fedfisher: Leveraging Fisher information for one-shot federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1612–1620. PMLR, 2024.

Marianne A Jonker, Hassan Pazira, and Anthony CC Coolen. Bayesian federated inference for estimating statistical models based on non-shared multicenter data sets. *Statistics in Medicine*, 43(12):2421–2438, 2024.

Marianne A Jonker, Hassan Pazira, and Anthony CC Coolen. Bayesian federated inference for regression models based on non-shared medical center data. *Research Synthesis Methods*, 16(2):383–423, 2025.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

Andrei Kazlouski, Ileana Montoya Perez, Faiza Noor, Mikael Högerman, Otto Ettala, Tapio Pahikkala, and Antti Airola. Towards practical federated learning and evaluation for medical prediction models. *International Journal of Medical Informatics*, 204:106046, 2025. doi: 10.1016/j.ijmedinf.2025.106046.

Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Proceedings of the 13th International Conference on Machine Learning (ICML)*, pages 284–292, 1996.

Ken Liu, Shengyuan Hu, Steven Z Wu, and Virginia Smith. On privacy and personalization in cross-silo federated learning. *Advances in neural information processing systems*, 35: 5925–5940, 2022.

Shuai Liu, Xiao Yan, Xiao Guo, Shun Qi, Huaning Wang, and Xiangyu Chang. Federated Bayesian network learning from multi-site data. *Journal of Biomedical Informatics*, 163: 104784, 2025.

Xiang Liu, Liangxi Liu, Feiyang Ye, Yunheng Shen, Xia Li, Linshan Jiang, and Jialin Li. FedLPA: One-shot federated learning with layer-wise posterior aggregation. *Advances in Neural Information Processing Systems*, 37:81510–81548, 2024.

Xu-Qing Liu and Xin-Sheng Liu. Markov blanket and Markov boundary of multiple variables. *Journal of Machine Learning Research*, 19(43):1–50, 2018.

Jianjie Ma and Krishnamoorthy Sivakumar. Privacy-preserving Bayesian network learning from heterogeneous distributed data. In *Proceedings of the 2006 International Conference on Data Mining (DMIN'06)*, pages 246–252, 2006.

Péter Marx, András Millinghoffer, Gabriella Juhász, and Péter Antal. Joint Bayesian modelling of internal dependencies and relevant multimorbidities of a heterogeneous disease. In *Conference on Probabilistic Graphical Models*, volume 52 of *Proceedings of Machine Learning Research*, pages 310–320. PMLR, 2016.

Csaba Nador, Andrea Valek, Attila Juhasz, Csilla Nagy, Eszter Bodrogi, Miklos Szabo, and Agnes Jermendy. Between-center variability in the outcome of VLBW infants is not affected by socioeconomic deprivation. *Pediatric Research*, 98:1323–1330, 2025. doi: 10.1038/s41390-025-03937-x.

Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*, 2013.

Beat Neuenschwander, Michael Branson, and David J Spiegelhalter. A note on the power prior. *Statistics in medicine*, 28(28):3562–3566, 2009.

Országos Kórházi Főigazgatóság (OKFŐ). *National in vitro fertilisation, obstetrics and perinatal registry: Hungarian perinatal registry data collection manual for neonatal intensive care units*. Országos Kórházi Főigazgatóság, February 2023. URL https://share.google/gRUK3JIkSXNV9NeI8. Accessed June 2, 2026.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101, 2024. doi: 10.1007/s10618-022-00867-8.

Elin Shaddox, Christine B. Peterson, Francesco C. Stingo, Nicola A. Hanania, Charmion Cruickshank-Quinn, Katerina Kechris, Russell Bowler, and Marina Vannucci. Bayesian inference of networks across multiple sample groups and data types. *Biostatistics*, 21(3): 561–576, 2020. doi: 10.1093/biostatistics/kxy078.

Shagun Sharma and Kalpna Guleria. A comprehensive review on federated learning based models for healthcare applications. *Artificial intelligence in medicine*, 146:102691, 2023.

Tomi Silander, Janne Leppä-Aho, Elias Jääsaari, and Teemu Roos. Quotient normalized maximum likelihood criterion for learning Bayesian network structures. In *International conference on artificial intelligence and statistics*, pages 948–957. PMLR, 2018.

Jacopo Talpini, Marco Savi, and Giovanni Neglia. FedBEns: One-shot federated learning based on Bayesian ensemble. *arXiv preprint arXiv:2503.15367*, 2025.

Dipti Theng and Kishor K Bhoyar. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, 66 (3):1575–1637, 2024.

I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy, filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342, 2003.

Margrietha van der Linde, Nèwel Salet, Nikki van Leeuwen, Hester F. Lingsma, and Frank Eijkenaar. Between-hospital variation in indicators of quality of care: a systematic review. *BMJ Quality & Safety*, 33(7):443–455, 2024. doi: 10.1136/bmjqs-2023-016726.

Stephen Walker and Nils Lid Hjort. On Bayesian consistency. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(4):811–821, 2001.

Xingyu Wu, Bingbing Jiang, Yan Zhong, and Huanhuan Chen. Multi-target Markov boundary discovery: Theory, algorithm, and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4964–4980, 2023. doi: 10.1109/TPAMI.2022.3199784.

## Appendix

**Algorithm 1:** LOCO-FL: leave-one-center-out federated inference via local subtraction.

**Input:** Per-center datasets $D_1, \ldots, D_C$; family set $\mathcal{F}$; BD prior pseudocount $\eta$; scaling weights $(\alpha, \beta)$.

**for** *each family* $(V, \mathrm{pa}_V) \in \mathcal{F}$ ***in parallel*** **do**
  **for** *each center* $c = 1, \ldots, C$ **do**
    Locally compute count tensor $N^{(c)}_{V,\mathrm{pa}_V}$ from $D_c$
  **end**
  Securely aggregate $N_{V,\mathrm{pa}_V} \leftarrow \sum_c N^{(c)}_{V,\mathrm{pa}_V}$ at the server
  Server broadcasts $N_{V,\mathrm{pa}_V}$ to all centers
  **for** *each center* $c = 1, \ldots, C$ ***in parallel*** **do**
    Compute LOCO count by subtraction: $N^{(-c)}_{V,\mathrm{pa}_V} \leftarrow N_{V,\mathrm{pa}_V} - N^{(c)}_{V,\mathrm{pa}_V}$
  **end**
**end**
**for** *each center* $c = 1, \ldots, C$ ***in parallel*** **do**
  Run ordering-MCMC on the LOCO posterior $p(G \mid D_{-c})$ using $\{N^{(-c)}_{V,\mathrm{pa}_V}\}_{(V,\mathrm{pa}_V)\in\mathcal{F}}$
  Run ordering-MCMC on the local posterior $p(G \mid D_c)$ using $\{N^{(c)}_{V,\mathrm{pa}_V}\}$ (or on the reweighted score (1) with weights $(\alpha, \beta)$)
  Compute per-center Jaccard distances between LOCO, local, and global posterior samples
**end**

**Output:** Per-center MBS strata, posterior spreads $\bar{J}$, and LOCO-vs-local distances.