FLAI Protocol: Privacy-Preserving Contribution-Aware Federated Intelligence

Balkeum Labs
contact@balkeumlabs.ai
www.balkeumlabs.ai

Overview

Critical datasets, such as hospital EHRs, biometric and wearable signals, financial logs, and operational records, remain locked inside institutional silos because sharing raw personal data is prohibited under regulations such as **HIPAA** (as well as GDPR and PIPA). At the same time, organizations require cross-silo collaboration to build predictive models and generate real-world evidence. Original data providers receive neither transparency nor compensation.

Balkeum Labs provides a new privacy-preserving AI infrastructure enabling collaborative training and analytics without ever transferring raw data. Two breakthroughs make this possible at scale:

- **Contribution-Aware Federated Learning/Analytics (FL/FA)** — A compliance-first computing paradigm where no personal data leaves the source, while contributors are rewarded based on the quality and impact of their datasets.
- **Automated On-Premise LLM Standardization Agent** — A local model that harmonizes heterogeneous data into standardized schemas required for FL/FA, eliminating the months-long preprocessing traditionally needed.

Together, these innovations allow institutions and individuals to collaborate safely, fairly, and instantly.

Key Innovations

1. Contribution-Aware Federated Learning and Federated Analytics

Balkeum Labs' core engine is a unified **Contribution-Aware FL/FA system**, enabling both predictive modeling and aggregated analytics without raw data ever leaving its source.

Why this matters for compliance:

- No identifiable EHR or personal data leaves devices or institutional servers.
- No cross-border data transfer ever occurs.
- No centralized storage of personal information.

This architecture is therefore **natively HIPAA-compliant**

Contribution awareness distinguishes Balkeum Labs' federated system from traditional approaches:

- Each contributor's update is evaluated and scored based on its actual positive impact on model accuracy, generalization, and representativeness.
- Rewards are distributed proportionally to contribution quality, not equally.
- High-value datasets (e.g., rare cohorts, clean longitudinal records, high-quality biometrics) earn significantly more.

This ensures the system is **capitalistic and merit-based**, in contrast to traditional FL systems where contributors receive equal influence regardless of contribution quality.

FL and FA run on the same privacy-preserving infrastructure, enabling both predictive and analytic collaboration simultaneously.

2. Automated On-Premise LLM Data Standardization

Before federated learning can occur, all participants must share the same data schema — historically the most expensive and time-consuming barrier to adoption. Balkeum Labs solves this with the **On-Premise LLM Standardization Agent**, which runs locally in a secure environment.

Capabilities:

- Detects and maps arbitrary schemas into unified FL/FA-ready formats
- Normalizes units, structures, and naming conventions
- Translates clinical codes (ICD, CPT, FHIR)
- Converts free-text notes into structured features
- Aligns data to standards such as OMOP, FHIR, or project-specific schemas

All processing happens on-premise — no raw data is uploaded or exposed.

Because the LLM improves itself using federated training (without seeing data), standardization becomes increasingly accurate over time.

Core Technologies (High-Level)

Contribution-Aware Federated Intelligence

A unified engine supporting both predictive modeling and statistical aggregation, without data sharing.

On-Premise LLM Standardization

Automates schema alignment locally, removing human preprocessing effort entirely.

Minimal Secure Aggregation

Only encrypted model updates and analytic summaries are transmitted; cryptographic details are intentionally abstracted for clarity in the litepaper.

B2B Enterprise Use Cases

Balkeum Labs enables cross-institution collaboration without any exchange of sensitive information, allowing regulated organizations to generate shared value while maintaining full compliance and control.

Hospitals

Key Benefits

- HIPAA-aligned collaboration with no raw EHR movement or data-transfer agreements.
- Instant onboarding via automated on-premise LLM standardization (FHIR, OMOP, ICD, CPT).
- Ability to jointly build predictive models across institutions without exposing patient data.
- Automated extraction of structured features from free-text clinical notes without sending raw text off-premise.

Representative Applications

- Clinical risk scoring (readmission, deterioration, complications).
- Treatment-response modeling across multi-hospital cohorts.
- Federated real-world evidence (RWE) generation for specific diseases.
- Longitudinal trend analysis across regions without merging EHR databases.

Pharmaceutical Companies

Key Benefits

- Access to diverse, multi-institution RWD/RWE without acquiring or storing patient-level data.
- Ability to run federated analytics directly on hospital EHRs and patient-contributed biometric/genomic data.
- Full privacy preservation enabling global collaboration without cross-border data transfer.
- Pay-per-query and pay-per-inference options to reduce dependence on expensive RWE vendors.

Representative Applications

- Drug-response prediction models using federated multi-center datasets.

- Safety signal detection using pooled analytics across hospitals.
- Post-market surveillance and label expansion studies.
- Synthetic cohort generation through federated analytics for trial planning.

Insurance Companies

Key Benefits

- Privacy-preserving access to enriched RWE that combines clinical data with biometric signals contributed by individuals.
- No exposure to identifiable PHI during model training or analytics.
- Ability to collaborate with hospitals and users without holding sensitive data.
- Fair and risk-aligned actuarial modeling using longitudinal, multimodal datasets.

Representative Applications

- Federated risk-adjustment models for reimbursement optimization.
- Early detection of chronic disease progression using biometric + EHR features.
- Fraud detection models trained across multiple insurers without sharing claimant data.
- Actuarial forecasting enhanced by multi-source longitudinal datasets.

B2C Consumer Health Use Case

Balkeum Labs supports a consumer-facing application where individuals can unify and contribute their personal health data.

Comprehensive Personal Dataset Integration

Users may store:

- EHR data (FHIR downloads)
- Biometric and wearable data (sleep, heart rate, oxygen, activity)
- Genomics and lifestyle logs

Healthcare significance: Hospitals cannot provide biometric or wearable datasets. By sourcing them directly from individuals, Balkeum Labs enables Real-World Data (RWD) that combines clinical + biometric signals — previously impossible using only institutional data.

User Benefits

- Personalized insights generated locally
- Significant rewards for high-quality longitudinal health data
- Full control over consent and deletion
- Raw data never leaves the device

Regulatory Alignment

Balkeum Labs' architecture is **natively HIPAA-compliant**:

- Raw identifiable data never leaves its source
- No cross-border transfers
- No centralization of personal data
- All data transformations happen on-premise

This inherently satisfies GDPR and PIPA as well, without additional reconfiguration.

Roadmap

2026

- Pilot deployments with hospital and research partners
- Consumer app beta with unified EHR + biometric integration
- Standardization Agent v1 deployment

2027

- Expansion to pharmaceutical, insurer, and financial partners
- Contribution-aware analytics expansion
- Standardization Agent v2 trained through federated learning

2028+

- Cross-industry scaled deployments
- Multi-national FL/FA networks
- Extended regulatory adaptation across jurisdictions

Vision

Balkeum Labs aims to make safe, fair, and privacy-preserving data collaboration the global default. Through contribution-aware federated intelligence, valuable datasets earn appropriate rewards rather than being treated equally regardless of quality. Through on-premise LLM standardization, the last major barrier to FL/FA adoption — data harmonization — disappears entirely.

This infrastructure unlocks a future where individuals control their data, institutions collaborate safely, and federated intelligence becomes the worldwide standard for machine learning in regulated environments.