# Document Metadata

author: jfraze@mycomp.org

title: Data Origin and Provenance for AI Systems

nist\_function: Map

priority\_phase: Must

last\_reviewed: 2025-06-24

status: draft

## **Purpose**

To ensure transparency and integrity of training, fine-tuning, and operational datasets used in AI and GenAI systems through documentation of data origin and lineage.

## **Scope**

Covers all structured and unstructured data used in GenAI model development, validation, deployment, or retraining.

## **Policy Statement**

Hospitals must maintain a complete record of data origin, provenance, and use limitations for all AI systems, especially those trained or operating on protected health information (PHI) or synthetic data.

## **Roles and Responsibilities**

• Data Stewards: Track lineage, consent, and retention of source data.

• ML Engineers: Document preprocessing and synthetic data creation.

• Privacy Officers: Validate de-identification, consent, and legal compliance.

## **Implementation Phases**

### **Must Do**

• Maintain a provenance log for all model training and tuning datasets (MAP-3.1).

• Restrict use of unverified third-party datasets (AI 600-1 §2.4.1).

• Ensure all PHI-eligible inputs have verified patient consent.

### **Should Do**

• Apply data quality scoring and completeness checks.

• Store data provenance using machine-readable metadata.

• Separate synthetic from real patient data with clear labels.

### **Recommended**

• Use cryptographic or blockchain-based lineage controls.

• Align with SBOM-style model artifact inventorying (SP 800-218A §3.3).

## **References**

• NIST AI RMF: MAP-3

• NIST AI 600-1: §2.4.1

• HIPAA Privacy Rule: §164.514

• FDA SaMD: GMLP 2024 Guidance

## **Review Cycle**

Every 12 months or upon a model update involving new training data.