# Document Metadata

author: jfraze@mycomp.org

title: Responsible Use of Synthetic Data in GenAI Systems

nist\_function: Map

priority\_phase: Should

last\_reviewed: 2025-06-24

status: draft

## **Purpose**

To govern the ethical and compliant use of synthetic data for training, fine-tuning, testing, and evaluating generative AI systems, particularly in contexts involving PHI, patient scenarios, or medical tasks.

## **Scope**

Applies to all hospital teams using synthetic data to train or evaluate GenAI tools, including LLMs, summarizers, dialogue agents, and structured data generators.

## **Policy Statement**

Synthetic data may be used to reduce risk, improve privacy, and support model generalization, but it must be generated, validated, and labeled responsibly. Misuse or over-reliance on synthetic data for critical clinical decisions is prohibited.

## **Roles and Responsibilities**

• Data Scientists: Tag synthetic datasets, document generation methods, and validate similarity to real distributions.

• Compliance & Privacy: Confirm synthetic data complies with HIPAA de-identification criteria and ONC interpretability standards.

• AI Risk Committee: Approve synthetic data use for clinical training or public sharing.

## **Implementation Phases**

### **Must Do**

• Label all synthetic datasets clearly and store separately from real PHI (AI RMF MAP-2.1).

• Confirm no re-identification risk using k-anonymity or other statistical tests (HIPAA §164.514(b)).

• Document source models, prompts, or rules used in data generation (AI 600-1 §2.2.1).

### **Should Do**

• Conduct outcome fidelity checks to ensure synthetic data do not introduce spurious correlations.

• Validate synthetic scenarios against clinical domain knowledge.

• Track model performance divergence when trained on synthetic vs. real data.

### **Recommended**

• Use synthetic data only as augmentation, not sole training input for safety-critical tasks.

• Report synthetic data usage in model cards or public documentation.