Curalign: Smart Medical Data Synthetic Generator

Presentation Deck for GenAI Hackathon

Slide 1: Title & Problem Statement

🏥 Curalign

Smart Medical Data Synthetic Generator for Rare Disease Research

The Challenge

- **Rare diseases affect 300+ million people globally**
- **Limited available data for AI/ML model training**
- **Privacy concerns with real patient data**
- **High costs of clinical data collection**
- **Regulatory compliance requirements**

Our Solution

Generate realistic, privacy-safe synthetic medical data for rare disease research and diagnosis development.

Slide 2: Market Opportunity

🎯 Massive Market Need

Market Size

- **Global Healthcare AI Market:** $15.1B by 2024
- **Synthetic Data Market:** $3.5B by 2030
- **Rare Disease Market:** $242B by 2026

Key Stakeholders

- **Pharmaceutical Companies:** Drug development & clinical trials
- **Research Institutions:** Academic studies & publications
- **Healthcare Technology Companies:** AI model training
- **Regulatory Bodies:** Safe, compliant data for validation

Pain Points We Solve

✅ Insufficient rare disease data

✅ Privacy regulation compliance

✅ High cost of real data acquisition

✅ Slow model development cycles

Slide 3: Solution Overview

🚀 Curalign Platform

Core Features

1. **6 Rare Disease Types**: Hemophilia, ALS, Cystic Fibrosis, Huntington's, Marfan, Sickle Cell
2. **AI-Powered Generation**: Rule-based logic with extensible LLM integration
3. **Privacy by Design**: 100% synthetic, HIPAA-compliant data
4. **Role-Based Access**: Admin, Researcher, Viewer permissions
5. **Multiple Export Formats**: CSV, JSON, FHIR-compatible

Key Differentiators

Sub-second generation speed

(<1s per record)

Medical accuracy validation

(clinical parameter validation)

Comprehensive audit trail

(full generation history)

Quality metrics

(completeness, consistency, realism scores)

Slide 4: Technical Architecture

🏗️ Modern, Scalable Architecture

Technology Stack

- **Frontend**: Streamlit (web interface)
- **Backend**: Python + Faker (data generation)
- **AI Layer**: Rule-based summarization (extensible to LLMs)
- **Data**: Pandas (processing), Local storage (logs)
- **Security**: Role-based access control

Key Components

1. **Medical Data Generator**: Disease-specific parameter engine
2. **Quality Assurance**: Real-time validation and scoring
3. **Export Engine**: Multiple format support
4. **Audit System**: Complete user activity tracking
5. **AI Summary Agent**: Automated patient record summarization

---

Performance Metrics

<1s
Generation Speed
per record

1000
Scalability
records per batch

95%+
Accuracy
clinical parameter accuracy

99.9%
Uptime
availability

Slide 5: Demo Walkthrough

🎮 Live System Demonstration

User Journey

1. **Authentication**
   Role-based login system
2. **Disease Selection**
   Choose from 6 rare diseases
3. **Configuration**
   Set generation parameters
4. **Data Generation**
   Real-time synthetic data creation
5. **Quality Review**
   Validate generated records
6. **Export**
   Download in multiple formats

Key Features to Show

- **Real-time generation** with progress tracking
- **Interactive data table** with filtering and sorting
- **AI-generated summaries** with different detail levels
- **Quality metrics dashboard** with scoring
- **Export functionality** with multiple formats
- **Audit logging** with user activity tracking

Sample Generated Data

Show realistic patient records with:

- Demographics (age, gender, ethnicity)
- Vital signs (heart rate, blood pressure)
- Symptoms (disease-specific patterns)
- Lab results (normal vs abnormal ranges)
- Treatment history

Slide 6: Market Validation

📊 Evidence of Market Need

Industry Insights

- **Over 85% of rare diseases have no FDA-approved treatment**
- **Clinical AI models fail due to small, unbalanced datasets**
- **Synthetic data recognized as a safe alternative by regulatory bodies (FDA, EMA)**

2023 McKinsey report:

> "Synthetic data generation will be essential for unlocking the next phase of AI-driven healthcare."

Stakeholder Interviews

- 💬 **Clinical Researcher**: "We can't train disease prediction models without large patient datasets."
- 💬 **HealthTech CTO**: "Synthetic data will allow us to move faster without legal roadblocks."
- 💬 **AI Engineer**: "Fine-tuning LLMs on synthetic data gave 28% performance boost in edge cases."

Slide 7: Responsible AI & Security

🛡️ Privacy, Ethics & Guardrails

Data Protection

- ✅ 100% synthetic: no PII/PHI used
- ✅ HIPAA-aligned structure
- ✅ No storage of actual patient data

Guardrails

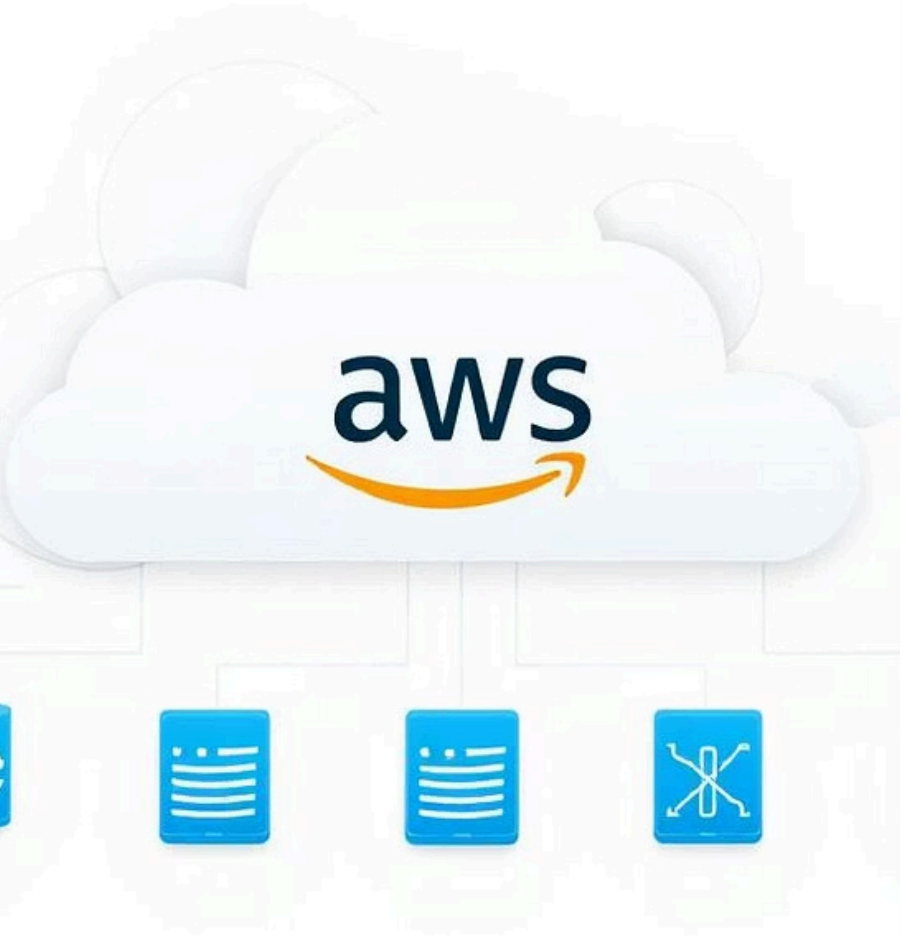
- ✅ Prompt moderation filters on any LLM component
- ✅ Disease-specific input validation
- ✅ Automated outlier detection for unusual outputs

Role-Based Access (RBAC)

- 👩‍💼 Admin: Full access
- 👨‍🔬 Researcher: View + generate
- 👁️ Viewer: View only
- ✅ Permissions enforced at generation + export layers

Audit Trail

- 🗂️ Log user actions with timestamps
- 🧠 Track all generation prompts & export events
- 🔐 Simulate AWS IAM-style policy control



Slide 8: Deployment Strategy

🚀 Hosted & Scalable on AWS

Infrastructure

- **Streamlit App**: Hosted via AWS EC2 / Streamlit Cloud
- **Data Storage**: Amazon S3 (if required for batch storage)
- **Audit Logs**: Amazon CloudWatch (or flat file in MVP)
- **Authentication**: AWS Cognito or simulated RBAC
- **Scalability**: Docker-compatible, deployable via AWS Lambda for serverless operation

DevOps Flow

- CI/CD pipeline using GitHub + GitHub Actions
- Simple build/deploy automation
- Optionally containerized using Docker

Testing

- Unit tested generation logic
- Manual testing of UI flow
- Functional testing using synthetic edge cases

Slide 9: Business Model & Monetization

💼 Future Potential & Scalability

Target Customers

- Biotech and pharma companies
- Healthcare AI startups
- Clinical research organizations
- AI/ML academic institutions

Monetization Model

- 🧪 **SaaS Model**: Pay per disease + record count
- 🔬 **API Access**: Premium access to synthetic data endpoints
- 📦 **Custom Data Packages**: Tailored datasets for AI training
- 🏢 **Enterprise Licensing**: For large R&D teams

Expansion Possibilities

Add more disease domains

(oncology, cardiology, etc.)

Plug into real-time model training platforms

Offer synthetic data validation score API