Culturally-Aware Synthetic Data Generation for Healthcare AI:

Ensuring Demographic Authenticity in Multicultural Contexts

Ameer Fahad Ali Khan

Oracle Corporation

ameerfahadali@gmail.com

September 2025

Keywords: *Synthetic Data, Healthcare AI, Cultural Sensitivity, Data Privacy, Name Generation, Demographic Authenticity, Southeast Asian Healthcare, AI Ethics*

Abstract

Synthetic data generation is essential for developing and testing healthcare AI systems without compromising patient privacy. However, most synthetic data tools produce culturally homogeneous outputs—typically Western names and demographics—that fail to represent the diversity of multicultural healthcare environments. This paper presents a methodology for generating culturally-aware synthetic healthcare data, demonstrated through the creation of 400,000+ patient records for a Southeast Asian hospital network serving four major ethnic communities. We detail the challenges of authentic name generation across distinct naming conventions, the importance of demographic proportionality, and the validation approaches ensuring statistical authenticity. The resulting dataset enabled realistic AI system testing while maintaining complete privacy compliance, offering a replicable framework for healthcare AI development in diverse cultural contexts.

1. Introduction

Healthcare AI systems require substantial datasets for development, testing, and validation. Using real patient data raises significant privacy concerns under regulations like HIPAA, GDPR, and regional equivalents [1]. Synthetic data generation offers a solution—creating realistic but fictional records that preserve statistical properties while containing no actual patient information.

However, existing synthetic data tools exhibit a critical limitation: cultural homogeneity. Popular libraries generate names like "John Smith" and "Jane Doe," demographics skewed toward Western populations, and patterns reflecting single-culture healthcare systems [2]. For healthcare organizations serving multicultural populations, such data fails to represent the diversity of actual patient populations.

This paper addresses this gap through a case study: generating 400,000+ synthetic patient records for a healthcare AI pilot [3] serving a multicultural Southeast Asian nation with four major ethnic communities, each with distinct naming conventions, demographic patterns, and healthcare utilization characteristics.

1.1 The Problem of Cultural Authenticity

Consider a healthcare AI system trained or tested on synthetic data containing only Western names. When deployed in a multicultural environment:

- User interfaces may not accommodate longer names or different character sets
- Search and matching algorithms may fail on unfamiliar name patterns
- Demographic analyses may not reflect actual population distributions
- Stakeholder demonstrations appear unrealistic and unrelatable

1.2 Contributions

This paper makes the following contributions:

1. A **methodology for culturally-aware synthetic data generation** applicable to multicultural contexts
2. **Detailed naming convention analysis** for four major Southeast Asian ethnic communities
3. **Validation framework** ensuring demographic authenticity
4. **Open discussion of ethical considerations** in culturally-sensitive data generation

2. Naming Convention Analysis

The target healthcare environment serves a multicultural nation with four major ethnic communities, each with fundamentally different naming conventions. Understanding these conventions was essential for generating authentic synthetic data.

2.1 Community A: Patronymic Naming System

The largest ethnic community uses a patronymic system where individuals are identified by their given name followed by their father's name, connected by "bin" (son of) or "binti" (daughter of).

- **Structure:** [Given Name] bin/binti [Father's Name]
- **Example (Male):** Ahmad bin Ibrahim
- **Example (Female):** Siti binti Abdullah
- **Key Challenge:** No family surname; father's name changes each generation

2.2 Community B: Family Name First System

The second-largest community traditionally places the family name before the given name, though Westernized formats also exist.

- **Traditional Structure:** [Family Name] [Given Name]
- **Example:** Tan Wei Ming, Lee Siew Ling
- **Key Challenge:** Limited set of common family names; given names often two characters

2.3 Community C: Given Name + Patronymic/Family System

The third community uses varied patterns, including standalone given names, given name with father's initial, or given name with family name.

- **Pattern 1:** [Given Name] s/o or d/o [Father's Name]
- **Pattern 2:** [Initial]. [Given Name]
- **Example:** Rajesh s/o Krishnan, K. Priya
- **Key Challenge:** Multiple valid formats within same community

2.4 Community D: Single Name or Informal Patterns

Migrant worker and indigenous populations sometimes use single names or informal naming patterns.

- **Pattern:** [Single Given Name] or [Given Name] [Village/Origin]
- **Example:** Suryadi, Dewi Lestari
- **Key Challenge:** No consistent surname; variable name lengths

Community	Name Structure	Has Surname?	Population %
Community A	Patronymic (bin/binti)	No	~55%
Community B	Family name first	Yes	~25%
Community C	Given + s/o, d/o, or initial	Sometimes	~10%
Community D	Single name / informal	Rare	~10%

Table 1: Naming Convention Comparison by Community

3. Generation Methodology

3.1 Name Pool Construction

For each community, we constructed pools of authentic names from public sources:

- **Given names:** 200+ names per community, gender-appropriate
- **Family names:** 50-100 common surnames (where applicable)
- **Honorifics:** Community-specific titles and prefixes

3.2 Generation Algorithm

```
def generate_patient_name(ethnicity, gender):
    if ethnicity == 'A':
        given = random.choice(COMMUNITY_A_GIVEN[gender])
        father = random.choice(COMMUNITY_A_GIVEN['male'])
        connector = 'bin' if gender == 'male' else 'binti'
        return f"{given} {connector} {father}"

    elif ethnicity == 'B':
        family = random.choice(COMMUNITY_B_FAMILY)
        given = random.choice(COMMUNITY_B_GIVEN[gender])
        return f"{family} {given}"
    # ... similar for communities C, D
```

3.3 Demographic Distribution

Patient records were generated with demographic distributions matching actual population statistics:

Attribute	Distribution	Source	Validated
Ethnicity	55/25/10/10%	Census data	✓
Gender	52/48% (F/M)	Healthcare utilization	✓
Age	Bimodal (pediatric + elderly peaks)	Hospital statistics	✓
Diagnosis	ICD-10 weighted by prevalence	National health data	✓

Table 2: Demographic Distribution Parameters

4. Validation Framework

4.1 Name Authenticity Validation

Generated names were validated through:

5. **Format compliance:** Names match expected patterns for each community
6. **Gender consistency:** Names appropriate for assigned gender
7. **Expert review:** Sample validated by cultural consultants
8. **Uniqueness check:** No accidental matches to real individuals

4.2 Statistical Validation

The generated dataset was validated against expected distributions:

- **Chi-square tests:** Ethnicity, gender, age distributions match census ($p > 0.05$)
- **Diagnosis patterns:** ICD-10 code frequency matches national health statistics
- **Geographic distribution:** Patient origins reflect facility catchment areas

5. Results

5.1 Dataset Statistics

Metric	Value
Total patient records generated	**400,000+**
Unique patient names	98.7%
Ethnic communities represented	4
Hospital facilities covered	15+
Name format compliance	**100%**
Real patient data used	**0 (fully synthetic)**

Table 3: Generated Dataset Statistics

5.2 Application in AI Pilot

The synthetic dataset enabled the healthcare AI pilot [3] to:

- Demonstrate realistic dashboards to healthcare executives
- Test natural language queries against diverse patient populations
- Validate UI handling of various name formats and lengths
- Ensure search/matching algorithms work across all naming conventions

6. Ethical Considerations

Generating culturally-specific synthetic data raises ethical considerations that must be addressed:

6.1 Avoiding Stereotypes

Care was taken to avoid encoding stereotypes in the data generation process. Diagnosis distributions were based on actual epidemiological data, not assumptions about ethnic groups. Income and insurance status were randomized independently of ethnicity.

6.2 Privacy of Cultural Knowledge

While naming conventions are public cultural knowledge, the specific combinations generated must never match real individuals. We implemented collision detection against public directories to ensure generated names do not accidentally identify real people.

6.3 Respect for Cultural Diversity

The goal of this work is to improve AI systems' ability to serve diverse populations accurately. By ensuring AI training and testing data reflects actual patient diversity, we aim to reduce algorithmic bias and improve healthcare equity.

7. Conclusion

This paper presented a methodology for generating culturally-aware synthetic healthcare data, addressing a critical gap in current synthetic data tools. Key contributions include:

- **Detailed naming convention analysis** for four distinct ethnic communities
- **400,000+ synthetic patient records** with 100% naming format compliance
- **Validation framework** ensuring demographic authenticity
- **Ethical guidelines** for culturally-sensitive data generation

The methodology is applicable to any multicultural healthcare environment, enabling AI development that respects and accurately represents diverse patient populations.

References

[1] Chen, R. J., et al. (2021). Synthetic data in machine learning for medicine. Nature Medicine, 27(12), 2082-2083.

[2] Ghorbani, A., & Zou, J. (2019). Data shapley: Equitable valuation of data for machine learning. ICML 2019.

[3] Khan, A. F. A. (2025). Agentic AI for Healthcare Operations: A Pilot Implementation of Natural Language Analytics Across Hospital Networks. SSRN Electronic Journal.

[4] Khan, A. F. A. (2025). Accurate Enterprise AI Analytics: A Reference Architecture for Anti-Hallucination SQL-RAG Systems on Oracle Cloud Infrastructure. SSRN Electronic Journal.

[5] El Emam, K., et al. (2020). Evaluating identity disclosure risk in synthetic health data. PLOS ONE, 15(3).

[6] World Health Organization. (2021). Ethics and governance of artificial intelligence for health. WHO guidance.