NHS Appointment No-Show Prediction – Enhanced Data Understanding Report

Overview

We created an enhanced synthetic UK NHS dataset to simulate patient appointment attendance behavior using real NHS statistical distributions. The dataset initially contained **250,000 records**, with **63,476 duplicates identified and removed**, resulting in **186,524 unique patient appointments**. Each record represents a patient appointment with various demographic, socioeconomic, and clinical factors derived from actual NHS Hospital Episode Statistics (2024-25).

Dataset Summary

Feature	Description
IMD_Decile	Index of Multiple Deprivation level (11 categories: 'Most deprived 10%' to 'Least deprived 10%', plus 'Unknown/Non-UK Country')
Ethnicity	18 ethnicity categories based on NHS ethnic codes including 'British (White)', 'Indian (Asian or Asian British)', and 'Unknown'
Age_Group	24 age ranges from '0' to '90-120', plus 'Unknown'
Gender	Male / Female
Medical_Specialty	82 NHS medical specialties (e.g., Cardiology, Midwifery, Ophthalmology, General Internal Medicine)
Consultation_Type	'Face-to-Face' or 'tele-consult' (Telemedicine)
Appointment_Type	'First' (initial appointment) or 'Subsequent' (follow-up)
Base_NoShow_Prob	Baseline demographic probability of no-show based on IMD and ethnicity (mean: 0.01)
NoShow_Prob_Final	Adjusted probability after combining all patient and appointment factors (mean: 0.04)
Previous_Appointments	Total number of past appointments (range: 0-14)
Previous_NoShows	Count of historically missed appointments (range: 0-4)
NoShow	Final outcome label ('Yes' / 'No')

Data Quality Metrics

- **Total Rows (Initial):** 250,000
- **Duplicates Found:** 63,476
- **Total Rows (After Cleaning):** 186,524
- **Missing Values:** None
- **Data Types:** 8 categorical, 2 numeric (float), 2 numeric (integer)

Basic Statistics

- **Average previous appointments:** 3.50 (SD: 1.87)
- **Average past no-shows:** 0.14 (SD: 0.38)
- **Average predicted no-show probability:** 0.04 (4.0%, SD: 0.023)
- **Overall no-show rate:** 5.26%

Demographic Analysis

Gender Distribution

Females constitute **56.4%** of the dataset compared to males at **43.6%**, reflecting typical NHS outpatient gender distributions where females tend to utilize healthcare services more frequently.

No-Show Rates by Gender:

- **Female:** 4.89% no-show rate
- **Male:** 5.73% no-show rate

Insight: Males show a notably higher tendency to miss appointments (+0.84 percentage points), consistent with NHS behavioral patterns where males generally have lower healthcare engagement.

Age Group Insights

The dataset shows an age distribution heavily weighted toward elderly patients, reflecting NHS outpatient demographics:

Top 5 Age Groups by Volume:

1. 80-84 years: 15,541 patients (8.33%)
2. 65-69 years: 14,577 patients (7.81%)
3. 70-74 years: 14,480 patients (7.76%)
4. 75-79 years: 14,435 patients (7.74%)
5. 60-64 years: 13,641 patients (7.31%)

No-Show Rates by Age Category:

Age Range	No-Show Rate	Interpretation
0-4 years	7.87-8.19%	Highest risk - dependent on parents/guardians
5-14 years	7.32-8.35%	High risk - young dependents
15-18 years	5.14-7.04%	Elevated risk - transitioning to independence
20-50 years	4.63-6.64%	Moderate risk - working age conflicts
60-84 years	4.56-4.98%	Lower risk - retired, health-conscious
85+ years	3.42-3.90%	Lowest risk - high healthcare engagement
Unknown	2.38%	Anomalously low (data quality issue)

Key Finding: Clear inverse relationship between age and no-show rate. Older patients (60+) are significantly more reliable attenders than younger cohorts, likely due to greater health awareness and fewer scheduling conflicts.

Socioeconomic (IMD Decile) Analysis

The Index of Multiple Deprivation distribution is relatively balanced across all deciles (~9-10% each), with a small proportion (1.84%) from unknown/non-UK locations.

No-Show Rates by Deprivation Level:

IMD Category	No-Show Rate
Most deprived 10%	5.29%
More deprived 10-20%	5.21%
More deprived 20-30%	5.29%
More deprived 30-40%	5.44%
More deprived 40-50%	5.22%
Less deprived 40-50%	5.34%
Less deprived 30-40%	5.32%
Less deprived 20-30%	5.17%
Less deprived 10-20%	5.20%
Least deprived 10%	5.17%
Unknown/Non-UK	4.74%

Insight: Minimal variation across deprivation levels (4.74-5.44% range). The expected gradient showing higher no-shows in deprived areas is very weak, with the most deprived 30-40% showing the highest rate (5.44%) but overall differences being minor. This suggests that **other factors (age, appointment type) may be stronger predictors** than socioeconomic status alone.

Clinical Factors Analysis

Consultation Type

The dataset reflects post-COVID NHS telemedicine adoption patterns:

- **Face-to-Face:** 82.6% of appointments (206,532 before cleaning)
- **Telemedicine:** 17.4% of appointments (43,468 before cleaning)

No-Show Rates:

- **Face-to-Face:** 5.88% no-show rate
- **Telemedicine:** 2.84% no-show rate

Key Finding: Telemedicine reduces no-show risk by **51.7%** compared to face-to-face appointments. This aligns with research showing remote consultations remove travel barriers and scheduling friction.

Appointment Type

- **First appointments:** 32.6% (81,469 before cleaning)
- **Subsequent appointments:** 67.4% (168,531 before cleaning)

No-Show Rates:

- **First:** 7.80% no-show rate
- **Subsequent:** 3.78% no-show rate

Key Finding: First appointments have **2.06× higher no-show risk** than follow-ups. This suggests:

1. Patients more committed once engaged in care pathway
2. Established relationships with healthcare providers improve attendance
3. Initial appointments may be speculative or lower priority

Medical Specialty Analysis

The dataset includes **82 distinct NHS medical specialties**, with the most common being:

Top 10 Specialty-Gender-Age Combinations:

Rank	Specialty	Gender	Age Group	Count
1	Midwifery	Female	35-39	1,166
2	Midwifery	Female	30-34	945
3	Obstetrics	Female	35-39	890
4	Midwifery	Female	40-44	842
5	Ophthalmology	Female	80-84	731
6	Obstetrics	Female	30-34	711
7	Allied Health Professional	Female	65-69	693
8	Gynaecology	Female	35-39	688
9	Allied Health Professional	Female	60-64	674
10	Ophthalmology	Female	85-89	667

Observations:

- **Reproductive health dominates** (Midwifery, Obstetrics, Gynaecology) for women aged 30-44
- **Ophthalmology** concentrated in elderly females (80+)
- **Allied Health Professionals** serve elderly populations (60+)

Most Common Specialties Overall:

1. Allied Health Professional: 29,911 patients (11.96% before cleaning)
2. General Surgery, Ophthalmology, Trauma & Orthopaedics: Combined ~15%

Ethnicity Distribution

Top 5 Ethnic Groups:

1. **British (White):** 66.87% (167,167 patients before cleaning)
2. **Indian (Asian or Asian British):** 6.33%
3. **Pakistani (Asian or Asian British):** 4.85%
4. **Any other White background:** 4.09%
5. **Bangladeshi (Asian or Asian British):** 3.34%

Data Quality Note Ethnicity capture appears much better with British (White) as the dominant known category (67%). Only 17 ethnic categories are represented (vs 18 initially), suggesting good data quality for demographic analysis.

Patient History Patterns

Previous Appointments

- **Distribution:** Poisson-like with $\lambda=3.5$
- **Range:** 0-17 appointments
- **Median:** 3 appointments
- **IQR:** 2-5 appointments

Previous No-Shows

- **Mean:** 0.14 no-shows per patient
- **Range:** 0-5 no-shows
- **75th percentile:** 0 (most patients have never missed)

Insight: The synthetic history generation successfully creates realistic patient profiles where:

- Most patients have limited appointment history (3-5 appointments)
- Historical no-shows are rare (85%+ have never missed)
- Chronic no-show offenders exist but are uncommon

Probability Distribution Analysis

Base_NoShow_Prob

- **Mean:** 0.01 (1%)
- **Standard Deviation:** 0.00
- **Range:** 0.01 to 0.01
- **All values identical**

Issue Identified: The base probability was clipped too aggressively during data generation, resulting in no variation. This limits the interpretability of demographic baseline risk.

NoShow_Prob_Final

- **Mean:** 0.040 (4.0%)
- **Standard Deviation:** 0.023 (2.3 percentage points)
- **Range:** 1.0% to 10.5%
- **Distribution:** Right-skewed with most patients at low risk

Insight: The final adjusted probability shows appropriate variation (1-10.5%) based on behavioral and appointment factors, even though the demographic baseline is uniform.

Key Findings & Insights

1. Overall No-Show Rate: 5.26%

- Realistic for UK NHS outpatient services (typically 4-8%)
- Within the expected range for NHS operations

2. Age is the Strongest Demographic Predictor

- Clear gradient: younger patients (0-18) show 2-3× higher no-show rates than elderly (85+)
- Effect size larger than gender or deprivation
- Peak no-show age: 5-9 years at 8.35%

3. Appointment Characteristics Matter More Than Demographics

- First appointments: 7.80% no-show (2× higher than subsequent)
- Telemedicine: 51.7% reduction in no-shows vs face-to-face
- These factors show stronger effects than IMD or ethnicity

4. Gender Gap is Moderate

- Males 0.84 percentage points higher across all age groups
- More pronounced than in previous dataset versions
- Consistent with healthcare engagement patterns

5. Deprivation Shows Weak Association

- Range: 4.74-5.44% across all deprivation levels
- Most deprived 30-40% shows highest rate (5.44%)
- Expected strong gradient is absent, suggesting other factors dominate

---

Class Balance Assessment

Target Variable (NoShow) Distribution:

- **No (attended):** 94.74% (176,717 patients after cleaning)
- **Yes (missed):** 5.26% (9,807 patients after cleaning)

Imbalance Ratio: 18:1 (No:Yes)

Implication: This is a **highly imbalanced classification problem**. Machine learning models will require:

- Class weighting or SMOTE oversampling
- Evaluation metrics beyond accuracy (F1-score, precision-recall, ROC-AUC)
- Stratified train-test splits to preserve minority class proportion

---

Data Readiness for Modeling

Strengths

1. **No missing values** - complete dataset after cleaning
2. **Duplicates removed** - 63,476 duplicate records cleaned, leaving 186,524 unique appointments
3. **Realistic distributions** - age, gender, specialty patterns match NHS data
4. **Rich feature set** - 12 features covering demographics, clinical, and behavioral factors
5. **Proper data types** - ready for encoding and analysis
6. **Good ethnicity capture** - 67% British (White), only 17 categories with minimal "Unknown" values

Limitations

1. **Base_NoShow_Prob uniformity** - limits demographic risk analysis (all values = 0.01)
2. **Class imbalance** - 18:1 ratio requires specialized handling in modeling
3. **Synthetic nature** - may not capture all real-world complexity and edge cases
4. **Duplicate rate** - 25.4% duplication rate (63,476 out of 250,000) suggests some feature combinations are over-represented

⍰ Data Quality Notes

- **Duplication pattern:** The high duplicate rate (25.4%) indicates certain demographic-clinical combinations were oversampled during synthetic generation
- **Impact on modeling:** Using cleaned dataset (186,524 rows) ensures each observation is unique and prevents overfitting to repeated patterns

Conclusion

The enhanced synthetic NHS dataset successfully replicates key patterns from real Hospital Episode Statistics:

- Age-stratified risk profiles
- Telemedicine adoption patterns
- Specialty-specific demographics
- Realistic appointment histories

After removing 63,476 duplicates, the dataset contains **186,524 unique patient appointments** with a **5.26% no-show rate** that aligns with NHS operational benchmarks. While the base demographic probability lacks variation, the final adjusted probabilities (1-10.5%) provide sufficient signal for predictive modeling.

The dataset demonstrates realistic clinical patterns including higher no-show rates for first appointments (7.80%), younger patients (8+%), and face-to-face consultations (5.88%). These patterns make the dataset suitable for testing intervention strategies that could help the NHS reduce missed appointment costs.

Key Takeaway: The data is clean, well-structured, and ready for feature engineering and machine learning model development.

Report Generated: October 30, 2025
Dataset Version: Enhanced Synthetic v1.0 (Cleaned)
Final Record Count: 186,524 unique appointments
Source: NHS Hospital Episode Statistics 2024–25 (synthetic reconstruction)

Prepared by: *Muhammad Aqeel*
LinkedIn: linkedin.com/in/aqeelkhan09
GitHub: github.com/maqeel019