**SYNTHETIC PATIENT DATA GENERATION DOCUMENTATION**

1. **OVERVIEW:**

This document outlines the methodology, logic, and constraints applied during the generation of synthetic patient data for a healthcare dataset using Python. The data generation process was guided by two key resources:

1. Table Artifacts.xlsx – This file defined the schema, data types, null constraints, format types, and statistical values (e.g., average, standard deviation) for each column.
2. Logical Artifacts.xlsx – This file provided grouped logical relationships and the required count of records for each combination of features like district, specialty, sub-specialty, doctor, procedure type, category, and age. *Note: The logical artifact content was authored by the user.*
3. **DATA SOURCES AND CONFIGURATION:**
4. Fake Data Library: Used Faker with the Australian locale (en\_AU) to generate location-specific synthetic data.
5. Files Used:

* Updated\_Logical\_Artifacts.xlsx: Provided base logical groupings and record counts.
* Latest\_Artifacts.xlsx (Table Artifacts sheet): Defined data rules for generation.

1. **KEY GENERATION RULES & CONSTRAINTS:**
2. **General Rules:**

* Each record is synthesized based on the constraints and structure defined in the artifacts.
* A patient\_counter ensures unique patient IDs.
* Null values are applied probabilistically using the null\_percentage from the Table Artifacts.

1. **Logical Constraints:**

* Logical combinations of ExtractDate, District, FacilityDesc, SpecialtyName, SubSpecialtyName, DoctorName, ProcedureType, Category, and Age were used to generate matching records.
* The count of records for each combination was exactly matched.

1. **Demographic Fields:**

* **PatientGNames, PatientSurname**: Generated using gender-specific names.
* **PatientGender and GenderCode**: Consistent, binary gender based on name.
* **PatientDoB**: Derived from Age to ensure internal consistency.

1. **Location Fields:**

* **PatientSuburb & PostCode**: Generated using Faker to reflect Australian regions.

1. **Categorical Fields:**

* **DoctorCode, SpecialtyCode, SubSpecialtyCode, Facility, Unit**: Consistent for identical names.
* **TheatreSpecialtyName**: Same as SubSpecialtyName.
* **ElectiveID**: Unique per sub-specialty and patient, formatted as EI-<SubSpecCode>-<PatientNumber>.

1. **Medical Fields:**

* **Procedure Mapping**: Based on ProcedureType, a consistent PrimaryProcedureDesc and PrimaryProcedureCode were assigned.
* **NRFCReason and NRFSComment**: Meaningful reasons and comments provided only for NRFC categories.
* **SurgeryReadyDate**: Randomized within a realistic window.

1. **Status Logic:**

* **CurrentStatusCode** is generated based on Category:
  + 'NRFC' → 'R' (Removed)
  + 'E', '1', '2', '3' → 'B' (Booked)
  + Else → 'W' (Waiting)
* **CurrentStatus** is mapped from the code using: { 'B': 'Booked', 'C': 'Cancelled', 'W': 'Waiting', 'R': 'Removed' }

1. **Operational Fields:**

* **ReadyForCareDate, OperationDate**: Plausible intervals based on surgery readiness.
* **NRFCDays and FutureNRFCDays**: Derived logically.
* **EstimatedLOS**: Always NULL as per requirement.
* **EstProcMins, SourceEstProcMin**: Positive values based on average and standard deviation.
* **WaitGroup**: Based on Category.
* **BookedBeyondBreach & Outsourcing**: Yes/No random values depending on logical status.

1. **OUTPUT:**

The generated data was written to **Synthetic\_Patient\_Data.xlsx** and contains all constraints and logical mappings as per the design.