This document provides a clear explanation of the Synthetic Hospital Readmission Dataset.

Dataset Overview

This synthetic dataset simulates hospital admission records[1]. Its purpose is to train and test an AI model that predicts the likelihood of a patient being readmitted to the hospital within 30 days[2]. Each row in the dataset represents a single hospital admission event[3].

Data Dictionary

The table below explains each column in the dataset[4].

Column Name	Description (Plain English)	Importance in a Hospital Setting	Typical Real-World Source
PatientID	A unique, random code for each patient[5555].	Allows patient records to be tracked without using personal details[6666].	Hospital EHR system[7].
Age	The patient's age in years[8].	A strong predictor of readmission risk, as older patients often have more complications[9].	Patient demographics database[10].
Gender	The patient's gender (Male, Female, Other)[11111111].	Can influence disease prevalence, treatment response, and	Patient registration system[13131313].

		readmission likelihood[12121212121212121212121212121212].	
AdmissionDate	The date the patient was admitted to the hospital[14].	Helps calculate the length of stay and track seasonal trends (e.g., winter flu admissions)[15].	Admission records[16].
DischargeDate	The date the patient was discharged from the hospital[17].	Used to calculate how long the patient was in the hospital[18].	Discharge summaries[19].
Diagnosis	The main medical condition treated during the admission (e.g., Heart Failure)[20].	Helps identify conditions with a high risk for readmission[21].	Clinical coding (ICD-10 codes)[22].
LengthOfStay	The number of days the patient was in the hospital[23].	Longer stays may suggest a more severe illness, which increases readmission risk[24].	Derived from admission & discharge dates[25].
PriorAdmissions	The number of times the patient has been admitted to the hospital before[26].	Frequent admissions are a strong indicator of chronic illness and higher readmission risk[27].	Hospital admission history[28].
Medications	Key medications prescribed during	Certain medications are	Pharmacy systems[31].

	the stay (e.g., Insulin, Antibiotics)[29].	associated with a higher likelihood of readmission[30].	
BMI	Body Mass Index, a measure of body weight relative to height[32].	A very high or low BMI can signal health risks that impact readmission[33].	Vital signs / nursing assessments[34].
SmokingStatus	The patient's smoking history (never smoked, former smoker, or current smoker)[35].	Smoking is a significant risk factor for many chronic diseases and complications[36].	Patient lifestyle history[37].
AlcoholUse	The patient's alcohol consumption (none, moderate, or heavy)[38].	Alcohol use can impact recovery, medication adherence, and the risk of readmission[39].	Patient lifestyle history[40].
BloodPressure	The patient's blood pressure upon admission (e.g., 130/85)[41].	High or low blood pressure is a key clinical indicator for many conditions[42].	Vital signs monitoring[43].
CholesterolLevel	The patient's cholesterol level in mg/dL[44].	High cholesterol is linked to cardiovascular diseases and an increased risk of readmission[45].	Laboratory results[46].

HbA1c	A blood test result showing the average blood sugar level over 3 months (%)[47].	Monitors diabetes control, which is a strong factor in readmission risk[48].	Laboratory results[49].
FollowUpAppointmentScheduled	Indicates if a follow-up appointment was scheduled after discharge (1 = Yes, 0 = No)[50].	Scheduling follow-up care helps ensure continuity of treatment and reduces readmission risk[51].	Discharge planning records[52].
InsuranceType	The patient's payment method (Public, Private, Self-Pay)[53].	Can influence a patient's access to care and follow-up services[54].	Billing / administrative systems[55].
ReadmittedWithin30Days	The outcome variable, indicating if the patient was readmitted within 30 days (1 = Yes, 0 = No)[56].	This is the target variable that the AI model is designed to predict[57].	Hospital readmission tracking[58].
RecordGeneratedAt	The exact date and time the synthetic data record was created[59].	Useful for tracking when the data was generated during testing[60].	Synthetic data generator[61].

Dataset Schema and Structure

The data fields are organized into logical groups to provide a complete view of a patient's

admission[62].

- **Demographics**: Basic information about the patient, such as age, gender, and insurance type[63].
- **Clinical Information**: Details about the hospital stay, including admission/discharge dates, diagnosis, treatments, and follow-up plans[64].
- **Lifestyle Factors**: Patient habits and body metrics that influence health, like smoking status, alcohol use, and BMI[65].
- **Lab & Vital Signs**: Objective clinical measurements taken during the hospital stay, including blood pressure and cholesterol levels[66].
- **Outcomes**: The target variable for prediction (readmission) and the record's creation timestamp[67].

Why This Data Matters for AI

The inclusion of detailed patient information makes the dataset more realistic and useful for AI model training[68].

- **Lifestyle Factors**: Fields like smoking, alcohol use, and BMI are strong predictors of chronic disease outcomes[69].
- **Lab Results**: Clinical data such as cholesterol, HbA1c, and blood pressure add clinical depth to the dataset[70].
- **Care Planning**: Information on follow-up appointments and insurance reflects real-world hospital processes that affect readmission risk[71].

Guide to Modifying Synthetic Data Generation

This table explains how to adjust the parameters used to generate the synthetic data[72].

To Change This...	Edit This Code...	Example Modification	Effect on Data
BMI range	bmi = round(random.uniform(18.0, 40.0), 1)	Change the 18.0 and 40.0 values.	Alters the weight distribution of the synthetic patient population[73].
Smoking status options	smoking_status = random.choice([...])	Add or remove items from the list, e.g., "Never", "Former", "Current".	Reflects different population habits[74].
Alcohol use options	alcohol_use = random.choice([...])	Add or remove items from the list, e.g., "None", "Moderate", "Heavy".	Adjusts the lifestyle risk factors in the dataset[75].
Blood pressure range	blood_pressure = f"{random.randint(100, 160)}/{random.randint(60, 100)}"	Change the number ranges for systolic (100-160) and diastolic (60-100) values.	Simulates patient populations with different blood pressure profiles[76].
Cholesterol range	cholesterol = random.randint(150, 300)	Adjust the minimum (150) and maximum (300) values.	Generates patients with higher or lower cholesterol levels[77].
HbA1c range	hba1c = round(random.uniform(4.5, 12.0), 1)	Adjust the 4.5 and 12.0 values.	Simulates different levels of diabetes severity and control[78].

Follow-up appointment rate	follow_up = random.choice([0, 1])	Change to a weighted choice to generate more 'Yes' or 'No' values.	Controls the likelihood of a follow-up appointment being scheduled[79].
Insurance types	insurance_type = random.choice([...])	Add or remove items from the list, e.g., "Public", "Private", "Self-Pay".	Reflects different healthcare payment systems[80].