I. Executive Summary

The following is a statistical analysis for 33,865 patients. The goal is to provide insights of two sub-groups (pre-auth and direct booking) patients including: appointments booked within 1-week, rescheduling rates, and gender and age distributions.

There is a total of 33,865 patients with 99.8% as pre-auth and 0.2% as direct booking. Overall 13% of pre-authorized patients book their appointment within one week of coverage. 15% of pre-authorized patients reschedule their appointment, with an average of 17 days, and 1% of direct booking patients reschedule their appointment, with an average of 1 day. Pre-authorized patients are 68% female, and 32% male. On average, 70% of each gender book an appointment. Majority of female pre-authorized patients (50%) are between ages 3 and 14, and 21 and 30. Majority of male pre-authorized patients (50%) are between ages 3 and 14, and 30 and 40.

Limitations of analysis and assumptions were made, to provide accurate statistical analyses.

II. Narrative

The following approach was carried out in order to analyze the data.

a. Understand the Data

   The data was sourced from external sources, and consists of two sheets: pre-authorized patients, and patients with appointments. “Pre-authorized patients” is a dataset consisting of patients that are pre-authorized to book an appointment. “Patients with appointments” is a dataset consisting of all patients (direct booking and pre-authorized) that have booked an appointment.

b. Acquire & Inspect the Data
   The following steps were carried out:
   i. Read data into script
   ii. Population or Sample Data? – for statistic purposes
   iii. df.head() – first 5 rows of dataset
   iv. df.info() – column names, data types, and number of non-null values
   v. df.shape()- number of rows and columns in dataset

   If the non-null value count does not equal the row count, this provides an insight into which columns have null/missing values.

c. Data Cleaning

The first step was to remove duplicate rows, and remove unnecessary columns. Checks were placed throughout the script, with command df.info(). The second step was to change data types of columns, and remove unnecessary characters and whitespaces if necessary.

i. Missing Data
    1. Pre-Auth: PATIENT_ID- structurally missing data, due to no PATIENT_ID if pre-authorized patient didn't book appointment
    2. Patients with Appointments: Source – missing at random

d. Exploratory Data Analysis

1. Determining distinct patient counts:

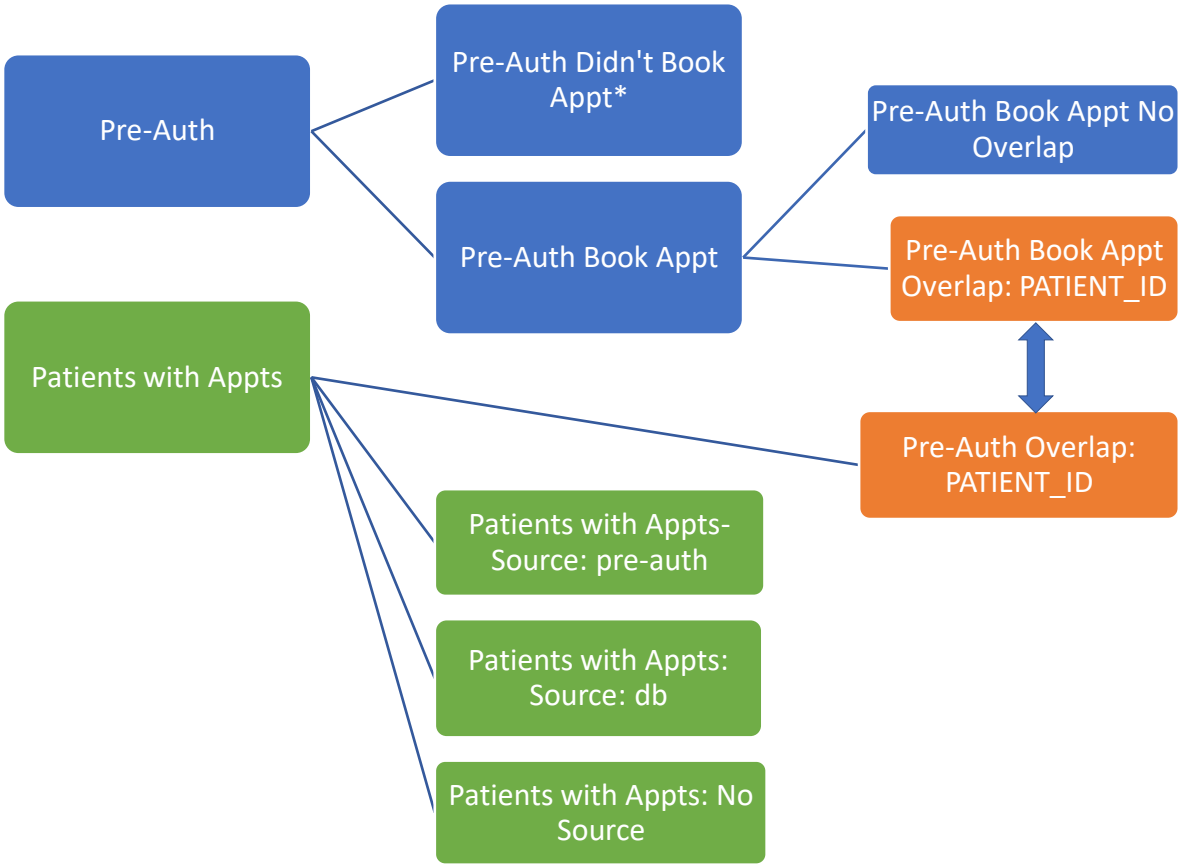

Figure 1. Data Map for Pre-Authorized and Patients with Appts Dataset

*no PATIENT_ID

i. Checked for no overlap in PATIENT_ID in sub tables.

ii. Issues Encountered When Creating Pre-Auth Book Appt Overlap Table on PATIENT_ID:

a. 12 PATIENT_ID duplicates in the Pre-Auth Book Appt sub-table. Further investigation was performed to determine if there was a pattern. It was determined all 12 duplicates have different PRE_AUTH_IDs and are unique observations. Having the same PATIENT_ID is an issue as this proves this is not a unique identifier for each patient. No PATIENT_ID duplicates in Patients with Appts dataset. Since the duplicate value was not a large part of the PATIENT_ID overlap

dataset (0.05%), it was determined to keep note of this and move forward. Data points for duplicate PATIENT IDs have same values for [‘FIRST_APPT_TIME’, ‘FIRST_NONCANCELLED_APPT_TIME’, ‘Source’].

b. Source values as ‘db’ in the PATIENT_ID Overlap sub-table. These Source values were replaced with ‘pre-auth’, and updated in Patients with Appts dataset.

2. Determining rescheduling rates:

ii. Pre-Auth Patients Issues:
   a. FIRST_NONCANCELLED_APPT_TIME values as ‘0’. These values were considered missing data and made up 10% of the dataset for pre-authorized patients. Since not enough data is known about the dataset to impute, these values were dropped.
      i. In order to drop missing data and have confidence in the conclusion, missing data should be 5% or less of the dataset.

iii. Direct Booking Patients Issues:
   a. FIRST_NONCANCELLED_APPT_TIME values as ‘0’. These values were considered missing data and made up 4% of the dataset for pre-authorized patients. Since not enough data is known about the dataset to impute, these values were dropped.
      i. In order to drop missing data and have confidence in the conclusion, missing data should be 5% or less of the dataset.

3. Focusing on gender distributions:
   Gender is another dimension I focused on. The following issue(s) were encountered.
   i. GENDER values as ‘0’. These values were considered missing data and made up 3% of the dataset for pre-authorized patients. Since not enough data is known about the dataset to impute, these values were dropped.
      1. In order to drop missing data and have confidence in the conclusion, missing data should be 5% or less of the dataset.

e. Data Analysis Results:

1. Distinct Count:

Pre-Authorized Patients Distinct Count	Direct Booking Patients Distinct Count	Total Patients Distinct Count
33784	80	33865

Figure 2. Patient Distinct Counts

There is a total of 33865 patients, with 99.8% as pre-authorized and 0.2% as direct booking. It can be seen there are significantly less direct booking patients than pre-authorized patients.

2. Rescheduling Rates:

Pre-Authorized Patients		
Patients with First Appointment within 1 week of coverage start (%)	Patients that Reschedule First Appointment (%)	Average Days between Rescheduled and First Appointment
12	15	17

Figure 3. Pre-Authorized Patient Appointments Statistics

Direct Booking Patients		
Patients with First Appointment within 1 week of coverage start (%)	Patients that Reschedule First Appointment (%)	Average Days between Rescheduled and First Appointment
N/A*	1	1

Figure 4. Direct Booking Patient Appointments Statistics

*see Assumptions

12% of pre-authorized patients book their appointment within 1 week of the coverage start date. 15% of pre-authorized patients reschedule their appointment, with an average of 17 days, and 1% of direct booking patients reschedule their appointment with an average of 1 day. There is a significant difference in rescheduling rate between pre-authorized patients versus direct booking patients (15% vs 1%), as well as in average days for rescheduling (17 days vs 1 day).

3. Gender and Age Distributions:

   Another dimension I chose to explore is Gender. Data for this variable was only present in the Pre-Auth dataset; the following analysis is for pre-authorized patients.

Female		Male	
% Female Patients	% Female that book appointments	% Male Patients	% Male that book appointments
68	73	32	67

Figure 5. Pre-Authorized Patients Gender % and % that Book Appointments

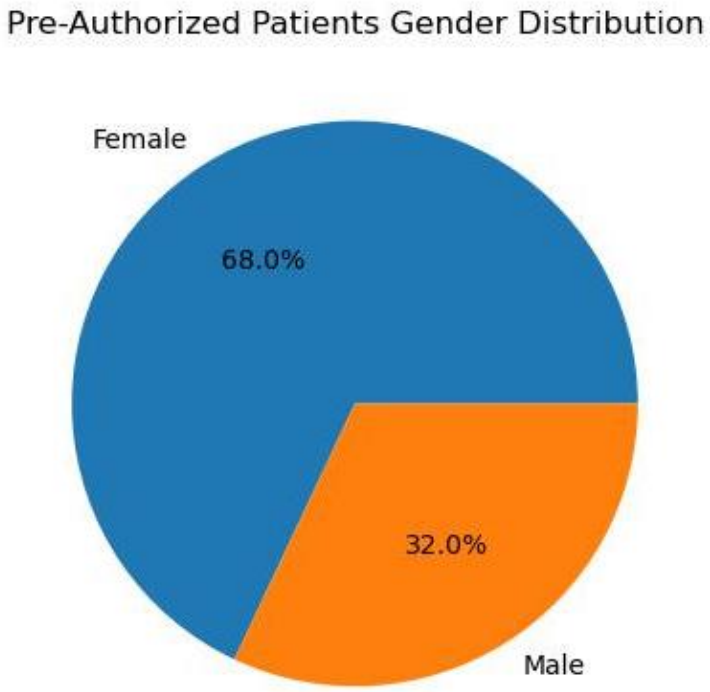

Figure 6. Pre-Authorized Patients Gender Distribution

The pre-authorized patients are 68% female and 32% male. Even though there is a significant difference in gender proportions, roughly similar proportions book appointments (~70%).

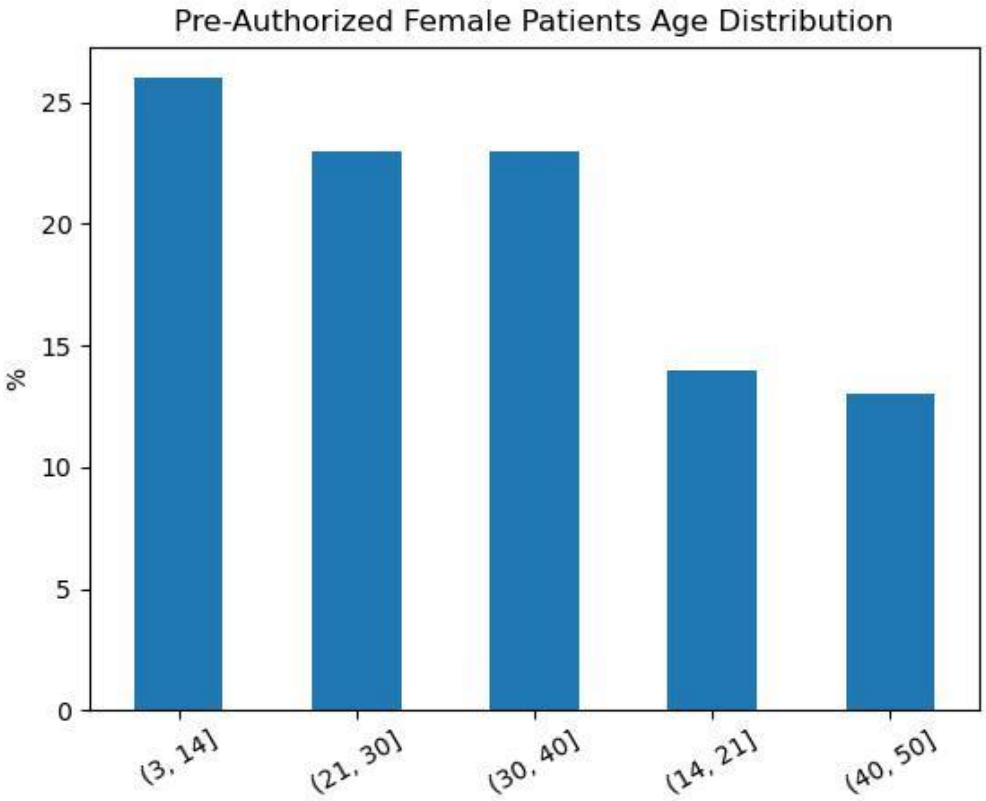

Figure 7. Pre-Authorized Female Patients Age Distribution

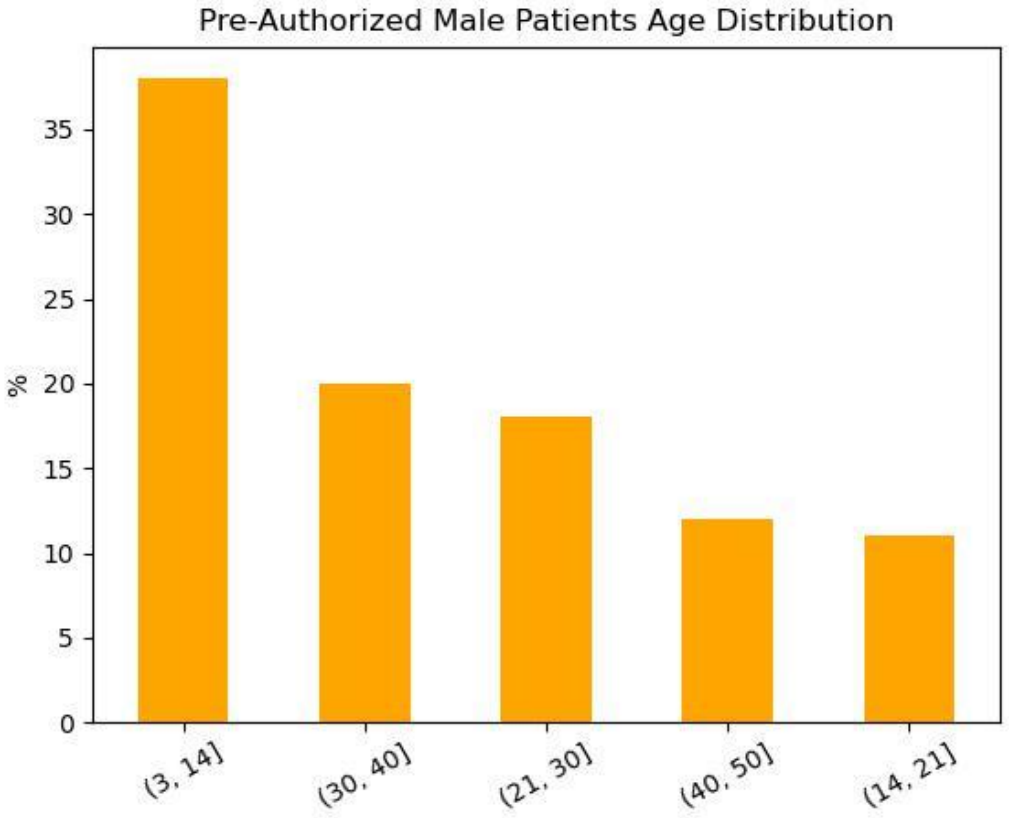

Figure 8. Pre-Authorized Male Patients Age Distribution

Figure 7 and 8 detail the age distribution for pre-authorized female and male patients, and are sorted from highest frequency to lowest frequency. For pre-authorized female patients, majority are between ages 3 and 14 (~25%), and 21 and 30 (~23%). For pre-authorized male patients, majority are between ages 3 and 14 (~35%), and 30 and 40(~20%).

III. Assumptions

The following assumptions were used in calculating statistics.

i. Overall:
   a. PRE_AUTH_ID is ONLY for pre authorized patients
   b. Only pre-authorized patients included in the pre-auth dataset
      i. This means if a PRE_AUTH patient is marked wrongly as 'db' in the Source column of the patients_with_appts dataset it is considered an entry error
   c. Pre-Auth patients who didn’t book an appt are not in patients with appts dataset

d. Direct Booking Patients are listed in Patients with Appts with source as 'db'

ii. Rescheduling Rates:
   a. Only considered Pre-authorized patients with appointments as base dataset

   b. Denominator only includes Pre-Authorized patients with COVERAGE_START_DATE < FIRST_APPT_TIME_START

   c. Rescheduling means FIRST_NONCANCELLED_APPT_TIME_START > FIRST_APPT_TIME_START

   d. Data is logged for FIRST_NONCANCELLED_APPT_TIME_START when patient attends first appointment

   e. If FIRST_NONCANCELLED_APPT_TIME_START == FIRST_APPT_TIME_START, it means no rescheduling

   f. If FIRST_NONCANCELLED_APPT_TIME_START < FIRST_APPT_TIME_START, included as no rescheduling.

iii. Gender Distribution:
   a. Only pre-authorized patients with ages greater than 3 were included in the age distribution analysis, since humans are not aware of oneself until age 3.

Notes:

iv. PATIENT_ID is the unique identifier for each patient in patients_with_appts dataset
v. Since no PATIENT_ID for pre-auth patients who didn’t book appt, unique identifier for each patient in the pre_auth dataset is PRE_AUTH_ID.