BIS568: Homework 1

Rong Sun

2025-02-06

1 Introduction

This report outlines the process of developing a cohort using SYNTHEA data structured in the OMOP (Observational Medical Outcomes Partnership) schema. The primary objective is to explore the data, perform necessary table merges, clean the dataset, and construct a structured and analyzable dataframe based on specified inclusion and exclusion criteria.

OMOP is a standardized data model designed for harmonizing diverse healthcare datasets, enabling interoperability across different sources. By leveraging this schema, we ensure consistency in data extraction and processing. The cohort is constructed by selecting living individuals aged 52–56 with recorded visits ranging from 1 to 37. To refine the dataset, we apply exclusion criteria, removing patients seen between visits 3–17 with high systolic blood pressure (the cutoff point for high systolic blood pressure is any value > 140) and those with visits lacking associated measurements.

To build the dataframe, we integrate key tables from the OMOP schema:

- **person:** Provides demographic details, including person_id, gender, race, ethnicity, and current age.
- **visit_occurrence:** Contains visit-related information, such as visit_date, age at visit, and visit length.
- **condition_occurrence:** Identifies the presence or absence of hypertension and diabetes.
- **measurement:** Includes all available biomarker measurements, with missing values recorded as NA.

After constructing the dataset, visualizations summarize key cohort characteristics, such as demographic distributions and visit patterns. Additionally, the *Short Answer Questions* section offers a deeper exploration of the tables and their implications.

2 Visualizing Descriptive Statistics of Cohort

To better understand the characteristics of the cohort, visualizations are used to summarize key demographic and visit-related variables.

2.1 Bar Plots (Figure 1)

Using R, bar plots are created to visualize the distributions of key demographic and visit-related variables, including patient age at visit, gender, race, and visit length, providing insights into the composition of the cohort and patterns in visit durations. These visualizations help identify any imbalances or dominant groups within the dataset.

From the bar plots generated for the variables:

- Patient Age at Visit: The age distribution is relatively uniform across the 1-year intervals from 52 to 56 years, indicating a balanced representation across these specific age groups.
- Gender: The gender distribution shows that approximately 60% of the cohort are males, highlighting a notable gender imbalance.
- Race: The race distribution reveals that around 80% of the cohort is White, with only about 10% Asian and 10% Black, underscoring limited racial diversity.
- Visit Length: Visit lengths are typically very short, with most visits lasting 0 days, though some extend up to 20 days.

Summary: The cohort has a balanced age distribution but is predominantly male, reflecting a gender imbalance. Racial diversity is limited, with most participants identifying as White. Additionally, most visits are brief, typically ranging from 0 to 20 days, with the majority lasting 0 days.

2.2 Box Plots (Figure 2)

Box plots are used to examine how current patient age varies across different factors, including visit number, gender, admission location, discharge location, race, and ethnicity. Presenting these plots as a subplot allows for comparison of patterns and differences across different categories, offering a clearer understanding of age distribution among patient subgroups and revealing potential trends and differences in patient characteristics.

From the box plots generated for current patient age by different variables:

In the upper-left plot, it can be observed that patients with a higher number of visits tend to be older, particularly those with six visits. In the top and bottom right plots, it is evident that Hispanic and male patients are generally younger, while female and non-Hispanic patients are older. Additionally, the bottom-left plot shows that Black patients are notably older than both Asian and White patients. The middle two plots provide no relevant insights regarding the locations where patients are admitted or discharged, meaning we cannot gain useful information about these aspects from them. (NA values in the plots may represent missing data or patients with no recorded information for certain variables.)

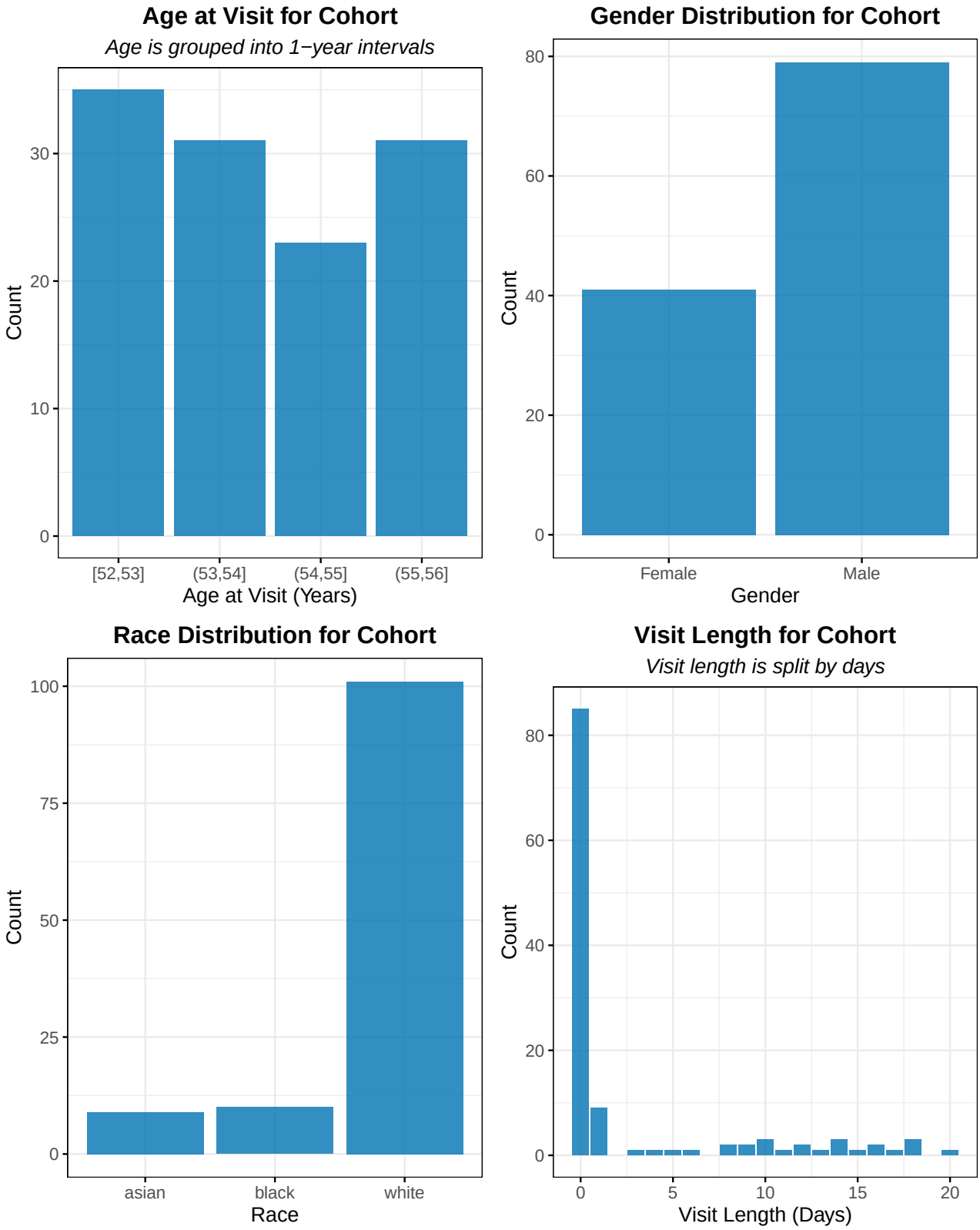

Figure 1: Bar plots for age, gender, race and visit length (days)

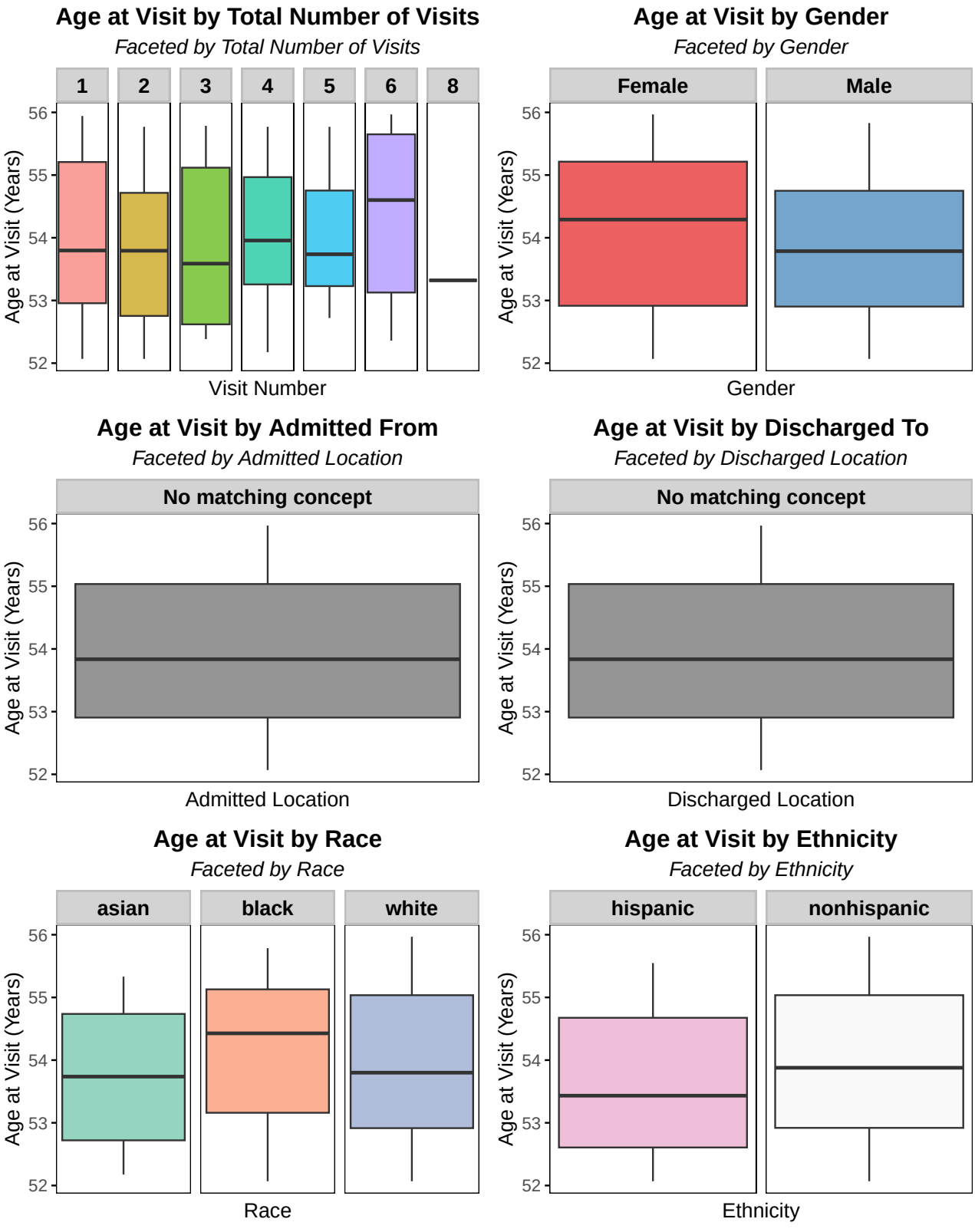

Figure 2: Box plot for age at visit between other variables

3 Short Answer Questions

3.1 Discussion of OMOP Tables and Their Relationship to Patient ID

The OMOP (Observational Medical Outcomes Partnership) Common Data Model (CDM) is a standardized data schema designed to harmonize healthcare data for analysis. Below is an explanation of the concept, person, and measurement tables and how they relate to the patient id:

(a) Concept Table

- Purpose: The concept table contains standardized medical concepts, such as diagnoses, procedures, drugs, and observations, mapped to standard vocabularies like SNOMED, LOINC, or RxNorm.
- Relationship to Patient ID: The concept table does not directly relate to the patient id. Instead, it serves as a reference table for standardizing terms used in other tables (e.g., measurement, condition_occurrence) to ensure consistent coding of medical events. For example, a concept_id in the measurement table might reference a specific lab test in the concept table.

(b) Person Table

- Purpose: The person table contains demographic information about each patient, such as birth date, gender, race, and ethnicity.
- Relationship to Patient ID: The person table is central to the OMOP schema, with each row representing a unique patient identified by person_id (equivalent to the patient ID). Serving as the primary key, person_id links each patient to their medical records across other tables, such as measurement and condition_occurrence, enabling the association of patient-specific data throughout the database.

(c) Measurement Table

- Purpose: The measurement table stores clinical measurements, such as lab results, vital signs, or other quantitative data.
- Relationship to Patient ID: The measurement table records clinical data, such as lab test results and vital signs, and links each measurement to a specific patient through the person_id column, which serves as a secondary key. The measurement_concept_id connects to the concept table, ensuring standardized terminology for each type of measurement. This structure allows for accurate and consistent classification of clinical measurements, such as blood pressure readings or cholesterol levels, with each entry tied to the appropriate patient using their person_id.

3.2 Feasible Question, Solution, and Machine Learning Approach

- Question: *Can we predict the risk of developing hypertension (high blood pressure) in patients based on their demographic information (e.g., age, gender) and historical lab measurements (e.g., cholesterol levels, BMI)?*
- Proposed Solution: To predict the risk of hypertension (high blood pressure) using OMOP data, we propose a solution leveraging demographic and clinical measurement data. First, we extract features such as age, gender, cholesterol levels, and BMI from the person and measurement tables, and hypertension diagnoses from the condition_occurrence table. We then use a **Random Forest classifier** to predict the binary outcome (hypertension: yes/no). This method is chosen because it handles both numerical and categorical data, captures non-linear relationships, and manages missing or imbalanced data effectively. The model's predictions can help identify high-risk patients early, enabling targeted preventive care and interventions.