Project 1

General Instructions

- Projects should be written in R and the report produced in RMarkdown (or Quarto if you already know & prefer it)
- Provide well-documented commentary on your code throughout your project & generally follow good programming practices
- Ensure your final report looks profession, produces well designed tables, etc. Tables should not go onto a second page unless that is unavoidable (long table).
- Hide your code in your final document
- Projects are due October 10, 2025 at 5pm EDT.

Background

Women with active ovarian cancer have chemotherapy appointments every two to three weeks. Physicians are concerned about patients visiting the emergency department (ED) between their chemo appointments and want to be able to predict at the end of the chemo appointment if the patient likely visit the ED or to be hospitalized (also called "Unanticipated hospital admission" or UHA) before the next appointment. The clinical team enlists the help of a biostatistician to create a predictive model. Before creating a predictive model, the data needs to be in a usable form. The goal of your project is to process the data to create a usable analytic dataset.

You can find the project data in the Project 1 assignment on Canvas. You will find two .csv files. One of these files contains *patient-level* data, i.e., data about the patients that are constant (each row is one patient). The other file is *encounter-level* data, i.e., data from office and hospital visits which varies over time (each row is one patient-encounter). An "encounter" is any contact a patient has with her medical care team. For this project, encounters are limited to office visits (chemo appointments), ED visits, and hospitalizations.

Patient data:

MRN	DOB	race	financialclass	ethnicity	hypertension	CHF	diabetes
DH1301	6/30/1957	Other	Private	non-Hispanic	N	N	N
JV9469	10/30/1964	White	Private	non-Hispanic	Y	N	N
TH8119	1/11/1981	White	Medicare	non-Hispanic	N	N	N
TJ3799	3/12/1949	White	Medicare	non-Hispanic	Y	N	Y
HP1319	9/3/1988	White	Private	non-Hispanic	Y	N	N

Encounter data:

MRN	contact_date	enc_type	temp	distress_score	WBC	BMI.r
HJ9754	6/26/2016	Office visit	97.91	2	15.12	28.33
GE5166	8/8/2016	Office visit	99.03	2	6.86	38.22
XV9573	1/20/2018	Office visit	99.15	2	5.48	32.13
CQ9338	7/5/2015	Office visit	99.09	3	15.11	25.09
DH1301	3/25/2018	Office visit	99.18	3	3.40	33.41

Project Components

Your project should include the following:

- Import both datasets. Do not hardcode a path!
- Merge the patient level data into the encounter level data
- This dataset will be your base "analytic dataset." Write a few sentences describing your dataset, including:
  - Granularity (what does each row represent?)
  - Dimensions (how many variables, how many rows?)
  - General description of variables (demographics, etc.)
- Address any data cleaning issues you see (missing data, implausible data, etc.)
  - Truncation for extreme values is preferable to setting to missing in this scenario
  - Note: You can consult with Brooke on what plausible ranges are for any variables you suspect to be implausible
- Re-categorize WBC (white blood cell count) into a categorical variable with the following levels:
  - Low (<3.2)
  - Normal (3.2-9.8)
  - High (>9.8)
  - Not Taken
- Create a table of the categorical WBC variable (Make sure order the table rows logically, not alphabetically). This table should include the counts (%) of encounters within each WBC group.
- Create & print a table of the counts & percentages of race, ethnicity, financial class, hypertension, congestive heart failure, and diabetes at the patient level. A table like this is often called a Table 1.