Project 1: Ovarian Cancer Analytic Dataset Preparation

Jiaqi Wang

2025-10-10

Introduction

Women with active ovarian cancer receive chemotherapy approximately every two to three weeks.Physicians are concerned about patients visiting the emergency department (ED) or being hospitalized between chemotherapy appointments.The goal of this project is to **process patient-level and encounter-level data** to create a clean, analytic dataset that will support future modeling of unanticipated hospital admissions (UHA).

1. Data Import

Both datasets are imported without hard-coding file paths using the here package.

2. Merge the patient-level data into the encounter-level data

After merging the patient-level and encounter-level datasets using **MRN** as the unique identifier, the analytic dataset now contains all encounter records with corresponding patient information.

Below is a brief preview showing the number of rows, variables, and the first few records.

```
[1] 550
```

```
[1] 550
```

```
 [1] "MRN"            "contact_date"   "enc_type"       "temp"          
 [5] "distress_score" "WBC"            "BMI.r"          "DOB"           
 [9] "race"           "financialclass" "ethnicity"      "hypertension"  
[13] "CHF"            "diabetes"      
```

Table 1: Preview of Analytic Dataset (first 10 ro

MRN	contact_date	enc_type	temp	distress_score	WBC	BMI.r	DOB	race
HJ9754	2016-06-26	Office visit	97.91	2	15.12	28.33	1999-06-05	White
GE5166	2016-08-08	Office visit	99.03	2	6.86	38.22	1993-09-16	White
XV9573	2018-01-20	Office visit	99.15	2	5.48	32.13	1976-09-27	White
CQ9338	2015-07-05	Office visit	99.09	3	15.11	25.09	1961-07-19	Black
DH1301	2018-03-25	Office visit	99.18	3	3.40	33.41	1957-06-30	Other
WQ8508	2019-08-25	Office visit	97.61	1	5.04	21.30	1970-05-16	White
XE4615	2017-06-20	Office visit	99.66	4	16.43	30.18	1997-02-11	Black
IO6623	2015-08-10	Office visit	99.43	2	2.87	26.04	1985-05-07	Other
JV9469	2014-04-11	ED/Hospitalization	98.32	NA	NA	-999.00	1964-10-30	White
NE9449	2019-02-15	Office visit	97.18	4	8.38	37.36	1966-07-20	White

3. Analytic Dataset Description

```
Granularity: One row represents one patient encounter.

Number of encounters: 550

Number of variables: 14

Unique patients: 50
```

The analytic dataset was created by merging the **encounter-level dataset** and the **patient-level dataset** using the variable *MRN* as a unique patient identifier.

Each row in this dataset represents a **single patient encounter**, which may correspond to an office visit, an emergency department (ED) visit, or a hospitalization.

The analytic dataset contains **550 encounters** from **50 unique patients** and includes **14 variables** in total.

The encounter-level variables capture clinical and visit-specific information such as contact date, encounter type, temperature, distress score, white blood cell count (WBC), and body mass index (BMI).

The patient-level variables include demographic characteristics (date of birth, race, ethnicity, and financial class) and comorbid conditions such as hypertension, congestive heart failure (CHF), and diabetes.