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.

Preview of Analytic Dataset (first 10 rows)

| MRN | contact\_date | enc\_type | temp | WBC | BMI.r | race | ethnicity | financialclass |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| HJ9754 | 2016-06-26 | Office visit | 97.91 | 15.12 | 28.33 | White | non-Hispanic | Private |
| GE5166 | 2016-08-08 | Office visit | 99.03 | 6.86 | 38.22 | White | non-Hispanic | Private |
| XV9573 | 2018-01-20 | Office visit | 99.15 | 5.48 | 32.13 | White | non-Hispanic | Private |
| CQ9338 | 2015-07-05 | Office visit | 99.09 | 15.11 | 25.09 | Black | non-Hispanic | Medicare |
| DH1301 | 2018-03-25 | Office visit | 99.18 | 3.40 | 33.41 | Other | non-Hispanic | Private |
| WQ8508 | 2019-08-25 | Office visit | 97.61 | 5.04 | 21.30 | White | non-Hispanic | Medicare |
| XE4615 | 2017-06-20 | Office visit | 99.66 | 16.43 | 30.18 | Black | non-Hispanic | Medicare |
| IO6623 | 2015-08-10 | Office visit | 99.43 | 2.87 | 26.04 | Other | non-Hispanic | Medicare |
| JV9469 | 2014-04-11 | ED/Hospitalization | 98.32 | NA | -999.00 | White | non-Hispanic | Private |
| NE9449 | 2019-02-15 | Office visit | 97.18 | 8.38 | 37.36 | White | non-Hispanic | Private |

## 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.

Together, these variables provide both longitudinal encounter data and baseline patient characteristics, forming a clean and well-structured analytic dataset that can be used to develop predictive models for unanticipated hospital admissions (UHA) among ovarian cancer patients.

## 4. Data cleaning

Rules applied (from project notes):

* **DOB:** unrealistic birth years set to missing (e.g., year < 1910 → NA).
* **BMI:** -999 is missing → recode to NA; then truncate to **10–50**.
* **WBC:** values < 0.05 treated as detection-limit error → set to **0.05**; truncate values > 50 to **50**.
* **Temperature / Distress:** constrained to plausible ranges (95–105 °F; 0–10).

After applying the cleaning rules above, we verified that implausible values were corrected.  
The table below compares the minimum and maximum values of key variables before and after cleaning.

Comparison of Selected Variables Before and After Data Cleaning

| Variable | Before\_Min | Before\_Max | After\_Min | After\_Max |
| --- | --- | --- | --- | --- |
| WBC | 0.00 | 53.60 | 0.05 | 50.0 |
| BMI.r | -999.00 | 352.13 | 10.00 | 50.0 |
| temp | 96.26 | 103.20 | 96.26 | 103.2 |
| distress\_score | 0.00 | 7.00 | 0.00 | 7.0 |

As shown above, implausible or out-of-range values were truncated to clinically reasonable limits, confirming that the dataset was successfully cleaned.

## 5. WBC Re-categorization

WBC is recategorized per assignment cut points.

## 6. WBC Summary Table

Counts and percentages of encounters within each WBC group.

| WBC\_cat | Count | Percent |
| --- | --- | --- |
| Low (<3.2) | 169 | 30.7% |
| Normal (3.2–9.8) | 196 | 35.6% |
| High (>9.8) | 113 | 20.5% |
| Not Taken | 72 | 13.1% |

## 7. Patient-Level Table 1

Baseline characteristics at the **patient level** (race, ethnicity, financial class, hypertension, CHF, diabetes).

| **Characteristic** | **N = 50**1 |
| --- | --- |
| **race** |  |
| Black | 10 (20%) |
| Other | 3 (6.0%) |
| White | 37 (74%) |
| **ethnicity** |  |
| Hispanic | 3 (6.0%) |
| non-Hispanic | 47 (94%) |
| **financialclass** |  |
| Medicare | 29 (58%) |
| Private | 21 (42%) |
| **hypertension** | 20 (40%) |
| **CHF** | 5 (10%) |
| **diabetes** | 2 (4.0%) |
| 1n (%) | |

## 8. Brief Summary

The report produced a single analytic dataset with encounter-level rows, merged with patient demographics and comorbidities.

Missing/implausible values were handled via explicit missing code conversion and clinically guided truncation.

WBC was categorized into Low, Normal, High, and Not Taken, and required summary tables were constructed.

This dataset is ready for downstream modeling of ED visits and unanticipated hospital admissions.