Project 1: Ovarian Cancer Analytic Dataset Preparation

Jiaqi Wang

2025-10-10

## 1. 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).

## 2. Data Import

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

## 3. Merge strategy and granularity

**Granularity:** one row per **encounter** (office visit, ED visit, or hospitalization).

We first do light, readable cleaning on each file, then left\_join() by MRN.

Analytic dataset: 550 rows x 14 columns.

## 4. Data cleaning

We apply rules:

* **DOB:** unrealistic birth years set to missing (e.g., year < 1910 → NA).
* **BMI:** -999 means missing → recode to NA; then truncate to **10–50**.
* **WBC:** values < 0.05 treated as detection-limit error → set to **0.05**; cap high values at **50**.
* keep temperature/distress in plausible ranges without dropping rows.

## 5. WBC Categorization

WBC is recategorized per assignment cut points.

## 6. WBC Summary Table

Counts and percentages of encounters within each WBC group.

Counts (%) of encounters within each WBC category

| WBC\_cat | n | 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).

Patient-level: Race

| race | n | % |
| --- | --- | --- |
| White | 37 | 74 |
| Black | 10 | 20 |
| Other | 3 | 6 |

Patient-level: Ethnicity

| ethnicity | n | % |
| --- | --- | --- |
| non-Hispanic | 47 | 94 |
| Hispanic | 3 | 6 |

Patient-level: Financial Class

| financialclass | n | % |
| --- | --- | --- |
| Medicare | 29 | 58 |
| Private | 21 | 42 |

Patient-level: Hypertension

| hypertension | n | % |
| --- | --- | --- |
| NA | 50 | 100 |

Patient-level: CHF

| CHF | n | % |
| --- | --- | --- |
| NA | 50 | 100 |

Patient-level: Diabetes

| diabetes | n | % |
| --- | --- | --- |
| NA | 50 | 100 |

## 8.Brief Summary

We 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.