Pre-Processing Description:

The code begins by loading multiple MIMIC clinical tables ( ADMISSIONS, DIAGNOSES_ICD, PATIENTS, PROCEDURES_ICD, and LABEVENTS). Disease cohorts are defined by matching ICD-9 diagnosis codes corresponding to the disease of interest. Patients containing at least one admission with a target ICD-9 code are classified into the disease cohort, while all remaining patients are assigned to the control cohort. Admission identifiers (HADM_ID) corresponding to disease diagnoses are extracted to enable event-level filtering and alignment of clinical history relative to disease onset.

At this point, it was noticed that there are patients in the admissions data with a written note describing AORTIC DISSECTION but who were not assigned one of the corresponding ICD-9 codes. Because we are not able to determine if those patients are or are not actual patients of interest, those patients are removed here.

Diagnoses and procedures are merged with admission, forming one row per admission capturing all sorted ICD-9 procedure and diagnosis codes. Diagnosis and procedure records are grouped by unique patient and admission identifier combinations and turned into list-based representations containing all diagnostic or procedural codes associated with each admission.

For each patient in the disease cohort, the admissions leading up and including their first diagnosis with the target disease are captured. A comparator is used to ensure admissions after their first diagnosis of at least one of the target conditions are not captured.

Admissions are ordered chronologically for each patient using admission timestamps. A patient-specific admission reverse index is generated ranking admission times per subject ID in descending order. Patients without disease onset timestamps are retained in the dataset without filtering.

All available lab events for each patient captured in the controls and diseased merged admissions dataframes are captured in two lab events dataframes, one for controls and one for diseased patients.

The final processed datasets contain admission-level structured clinical records for both disease and control cohorts. Each row represents a single hospital admission and contains aggregated diagnoses, procedures, demographic features, and temporal ordering information.