Data Wrangling Assignment – 3 (150 Points)

In this assignment we will use basic ideas of Set Theory to wrangle a clinical dataset using a programming language of your choice. Submit your code and datasets using your GitHub repository.

Here is a quick refresher to Set Theory: https://www.geeksforgeeks.org/set-theory/. These concepts are used to define inclusion and exclusion criteria of cohorts and in creation of their corresponding datasets (https://pubmed.ncbi.nlm.nih.gov/24026307/).

COVID-19 has been associated with the occurrence new diabetes and hyperglycemia (https://academic.oup.com/jamiaopen/article/4/3/ooab063/6320067). In this exercise you will use a synthetic diagnosis file containing patient IDs, ICD 10 diagnosis codes, and a date of diagnosis. You will need to use the following code sets for your wrangling steps:

- Diabetes Codes

ICD 10 Code	Concept
E08	Diabetes mellitus due to underlying condition
E09	Drug or chemical induced diabetes mellitus
E10	Type 1 diabetes mellitus
E11	Type 2 diabetes mellitus
E13	Other specified diabetes mellitus

- COVID Codes

ICD 10 Code	Concept
U07.1	COVID-19
J12.82	Pneumonia due to COVID-19

Questions:

1. Diabetes Set: (20 Points)
   a. Find all patients with Diabetes using the codes above by listing their patient IDs.
   b. Find the cardinality of the Diabetes set.
2. COVID Set: (20 Points)
   a. Find all patients with COVID using the codes above by listing their patient IDs.
   b. Find the cardinality of the COVID set.
3. Intersection Set (20 Points)
   a. Find all patients with Diabetes <u>and</u> COVID using the codes above by listing their patient IDs.
   b. Find the cardinality of the Intersection set.
4. Union Set (20 Points)
   a. Find all patients with Diabetes <u>or</u> COVID using the codes above by listing their patient IDs.
   b. Find the cardinality of the Intersection set.

5. Draw a Venn diagram showing the Diabetes, COVID, Intersection and Union sets. You might need to use a package. (40 points)
6. Diabetes only after COVID Set (30 points)
    a. Now including the date of diagnosis, find all patients with Diabetes only after they had COVID by listing their patient IDs.
    b. Find the cardinality of the Diabetes only after COVID set.
    c. Provide a count breakdown for each of the diabetes codes listed above occurring only after COVID.