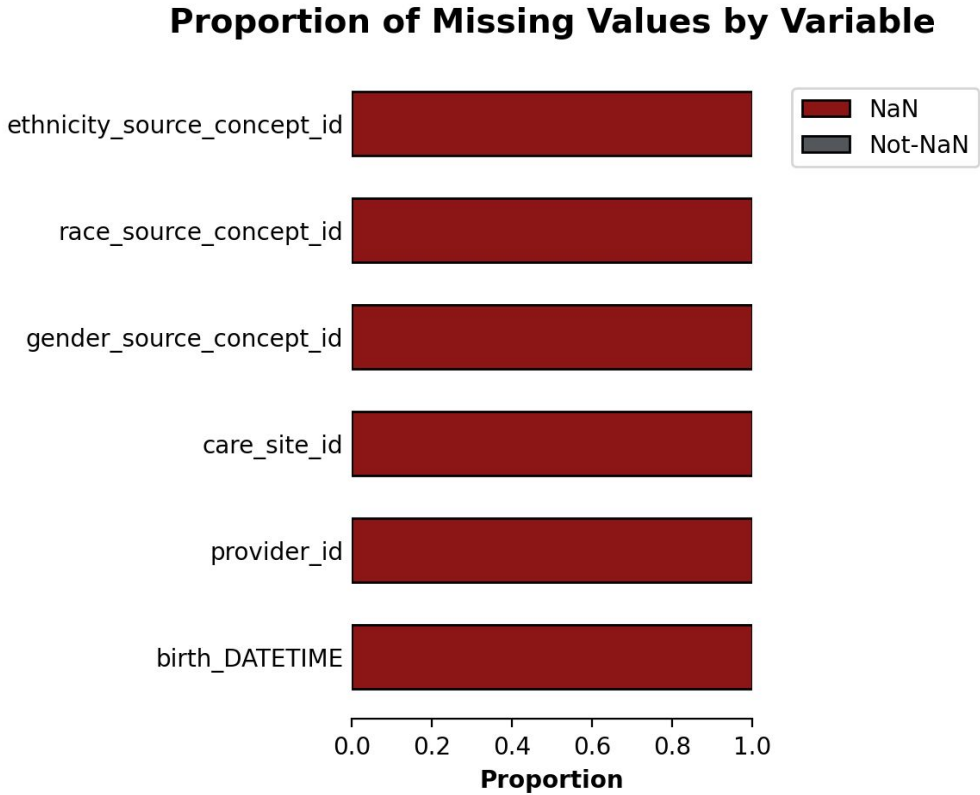
EHR Data Analysis: Biomedical Informatics

Samuel Campione

Stanford
MEDICINE

1a. Which variables in the person table have missing values? What number and proportion of those variables are missing?

- After inspecting the person table, I see that the following 6 variables have missing values:
  - birth_DATETIME
  - provider_id
  - care_site_id
  - gender_source_concept_id
  - race_source_concept_id
  - ethnicity_source_concept_id
- Each of these 6 variables is completely missing for all 1000 rows (all entries are null).
- I also checked the unique values in case NA values were assigned to non-standard codes for any variables (e.g., birth year of 0 instead NaN).



1b. Which variables in the person table have suspicious-looking values? (Choose 3 values to list, & explain why they' not what you expect).

- **Some variables have only 1 unique value.** Every patient's month_of_birth and day_of_birth is reported as "1". This is suspicious because it is unlikely that everyone in the sample has the same birth month and day. Further, all patients are categorized as 'White' with an ethnicity of 'Not Hispanic or Latino.' For a sample of 1000, it is unlikely that everyone is of the same race and ethnicity.
- **Birth year data is suspiciously distributed.** More than 60% of the patients are born before the year 1919. This is suspicious because the dates in the conditions occurrence and drug exposure tables are between 2007 and 2010, making a majority of the patients very elderly (some older than 100). Further, there are no patients born between 1919 and 1963. This is unexpected because it suggests there are no middle aged patients in this dataset, which raises questions about completeness and accuracy.
- **Gender is reported as a binary variable.** The variable gender_concept_id has only two unique values (Male and Female), but there is likely to be transgender people within a sample of 1000. This is suspicious as it may be unrepresentative of the population. It may also be dangerous to exclude this information because transgender patients may require special consideration, for instance, when being given a drug that could alter hormone levels.

2a. What is the concept ID for Congestive Heart Failure?

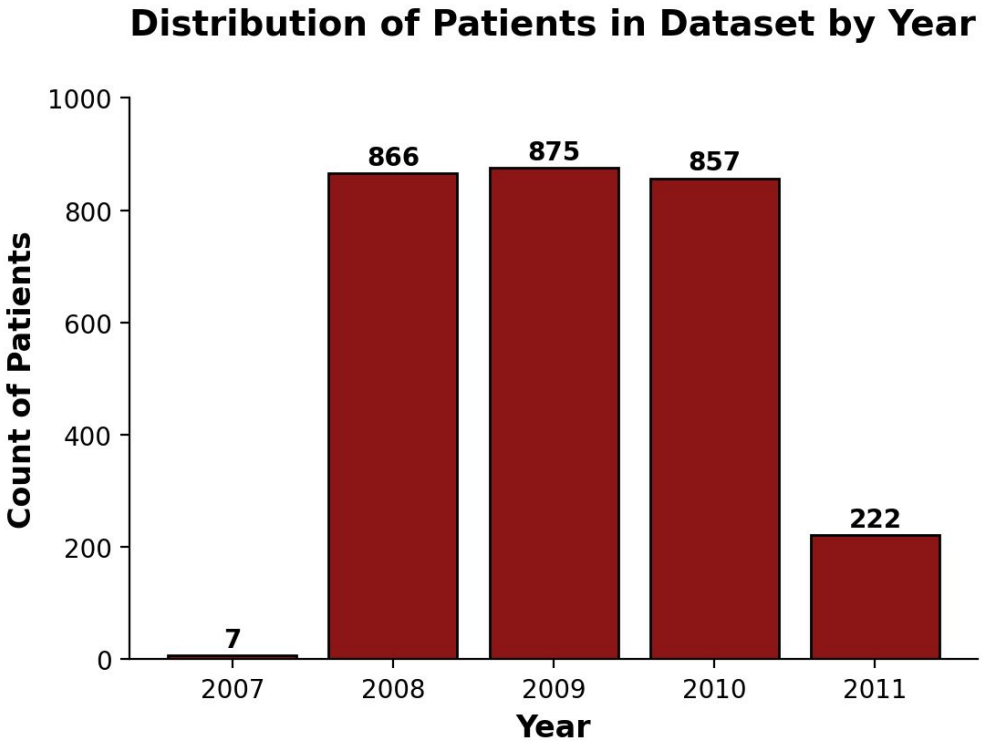
- The concept ID for Congestive Heart Failure is 319835.
- This concept ID corresponds to the SNOMED code for "Congestive heart failure", which specifically and clearly identifies Congestive Heart Failure without additional comorbid conditions or unspecified details—making it a precise and reliable choice for data analysis and machine learning.
- The concept table contains conditions comorbid with Congestive Heart Failure such as "Hypertensive heart disease with congestive heart failure" (concept ID 314378), "Benign hypertensive heart disease with congestive heart failure" (concept ID 312338), and "Malignant hypertensive heart disease with congestive heart failure" (concept ID 316994). The concept table also includes "Congestive heart failure, unspecified" (concept ID 44826642) which has a concept_class_id of '4-dig billing code' and vocabulary_id of 'ICD9CM', while the previously mentioned concepts have vocabulary_id of 'SNOMED'. ICD-9-CM codes are often used for billing and may not capture all clinical nuances.

2b. What clinical concept is represented by the concept ID of 8507?

- When I query the concept table where "concept_id = 8507", I observe that concept ID corresponds with the gender concept name 'MALE'.

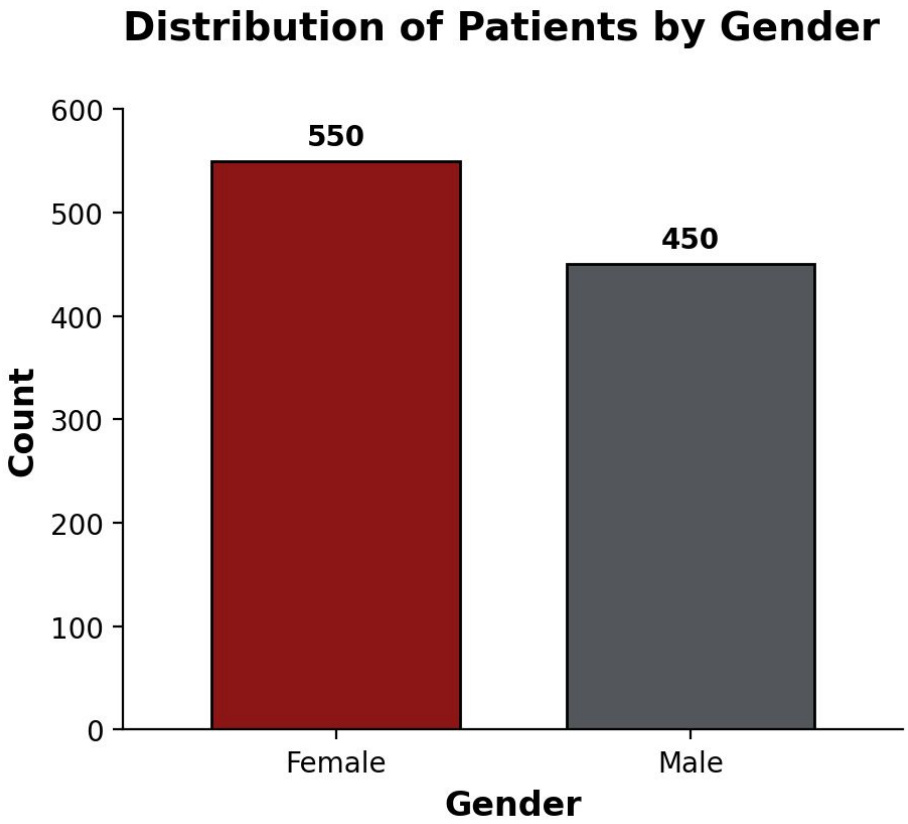
3a. How many patients are there for each year included in the dataset?

- To answer the question, I report the number of patients per year as the number of unique person_id's grouped by year in any of the condition_occurrence, drug_exposure, and death tables.



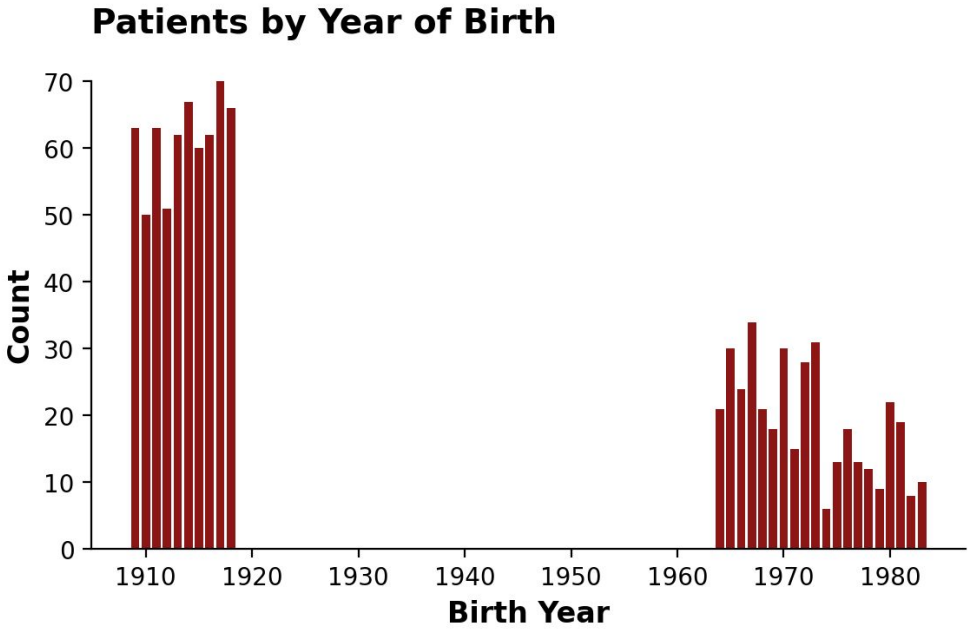
3b. What are the distributions of patients by race, ethnicity, and gender?

- Race has only one unique value, 'White'. Ethnicity also only has one unique value, 'Not Hispanic or Latino'. As stated earlier, this uniform distribution is likely due to an error in the dataset.
- Gender is distributed with 550 female and 450 male patients.



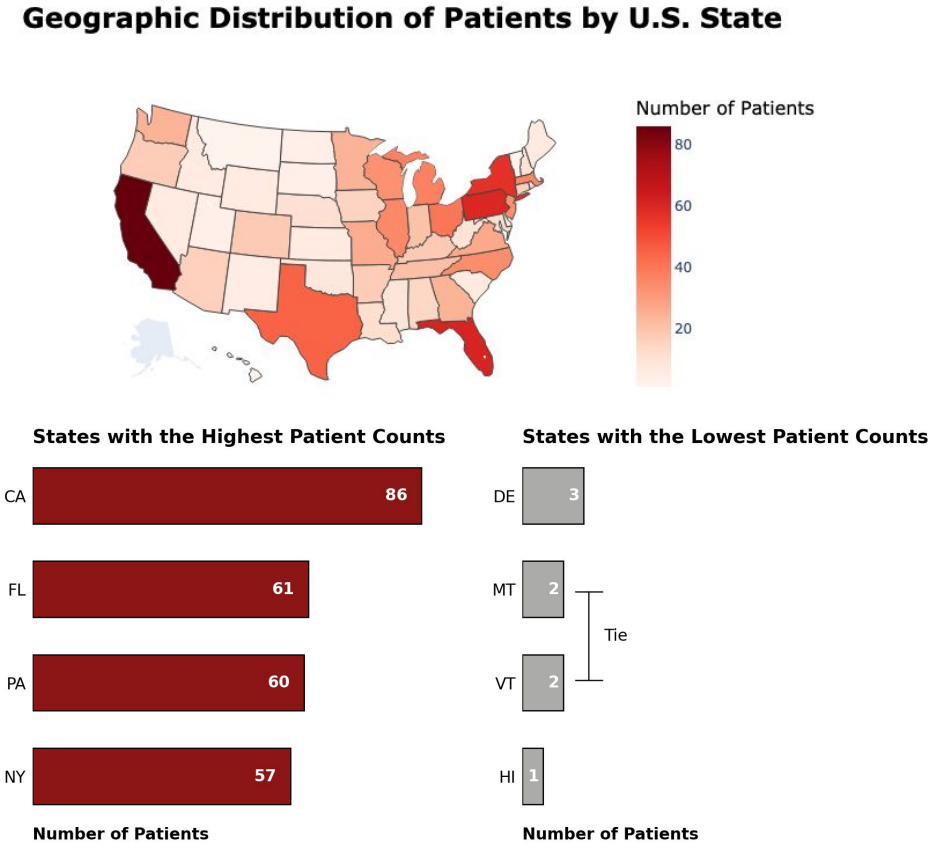
3c. What is the distribution of patients by year of birth and month of birth?

- All the patients in the dataset have a birth month of January, so the distribution of birth month is uniform. This is highly unlikely and indicates a potential data entry error or an issue with the data source.
- The distribution of birth years is bimodal with two distinct clusters of younger and older patients. The large gap between 1919 and 1963 raises questions of dataset completeness—assuming this was not an intentional choice in cohort selection. With a more complete dataset, I would hypothesize the distribution to be right-skewed as there are notably more older patients.



3d. What is the distribution of patients in different states? Which states have the most patients? And the least?

- The heatmap suggests coastal and northeast states have the most patients while midwest and western states have fewer.
- States with most patients:
  - California (86)
  - Florida (61)
  - Pennsylvania (60)
  - New York (57)
- States with least patients:
  - Hawaii (1)
  - Vermont (2) tied
  - Montana (2) tied
  - Delaware (3)
- This data contains a state labeled as '54' and is also missing the state Alaska. Perhaps Alaska was miscoded as '54'. In practice, I would inquire with the data source to gain more insight.

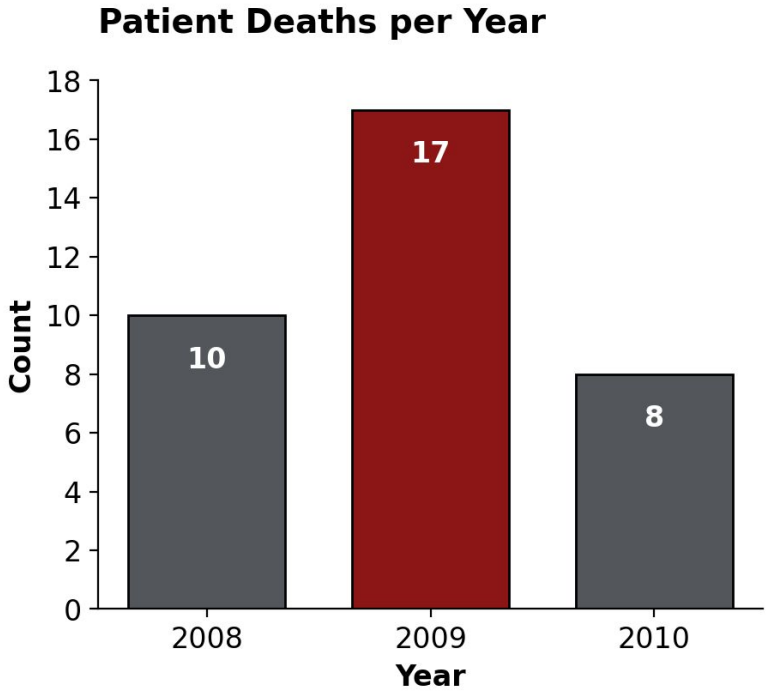


3e. How many of these patients have death data? Which state had the most deaths? Are there any states that saw no deaths?

- 35 patients have death data.
- Florida had the most deaths (4).
- There are 26 states in the dataset with no deaths.
- We can not make any inferences regarding Alaska as it is not explicitly included in the location table.

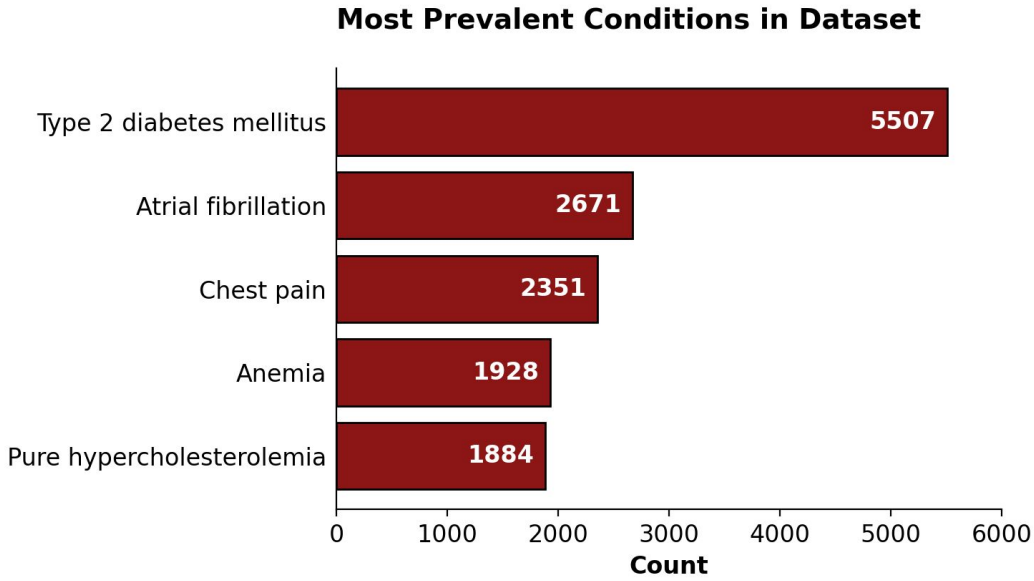
3f. Which year saw the most deaths?

- The year 2009 saw the most deaths (17).



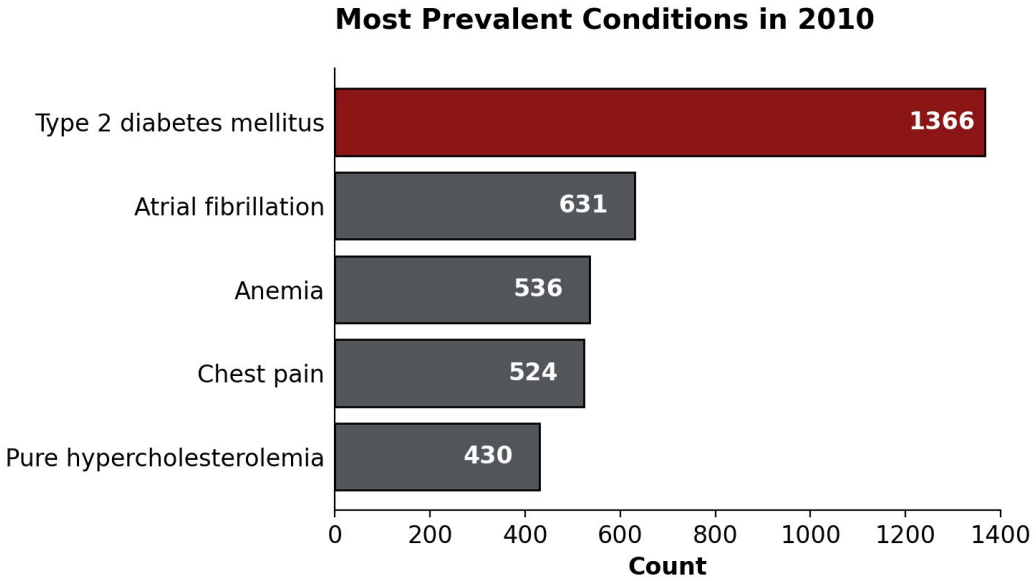
4a. What are the 5 most prevalent conditions in this dataset?

- The five most prevalent conditions in the dataset are Type 2 diabetes mellitus, Atrial fibrillation, Chest pain, Anemia, and Pure hypercholesterolemia.



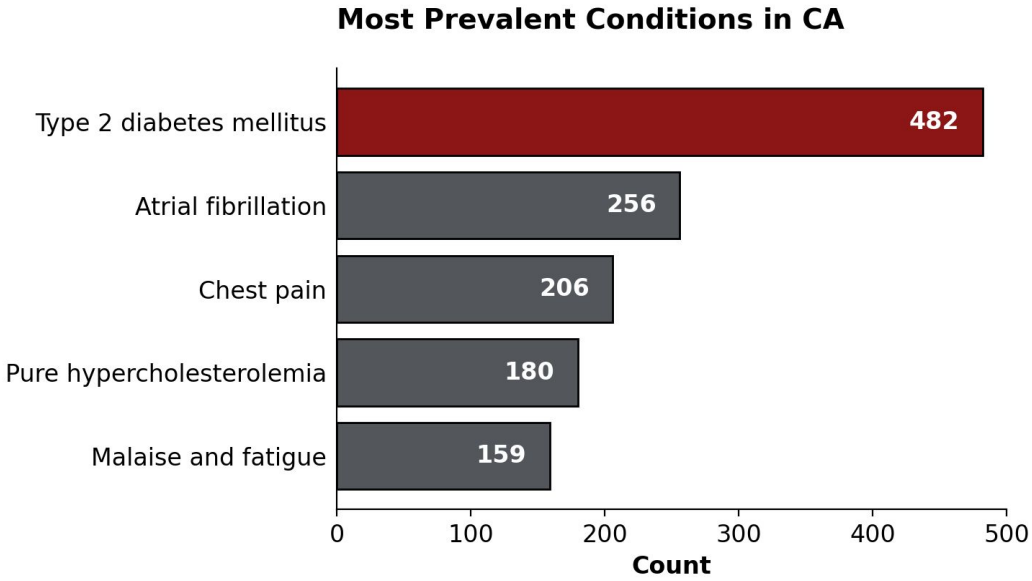
4b. Which condition was most prevalent in 2010?

- The top condition in 2010 was Type 2 diabetes mellitus with 1366 occurrences.
- Here, I assume prevalence in 2010 indicates the condition start date is in 2010, the condition end date is in 2010, or the year 2010 falls between the condition start date and end date.



4c. What was the most prevalent condition in CA (looking across all years)?

- Type 2 diabetes mellitus was the most prevalent condition in CA across all years with 482 occurrences.



4c. Which state had the highest incidence of Congestive Heart Failure across all years?

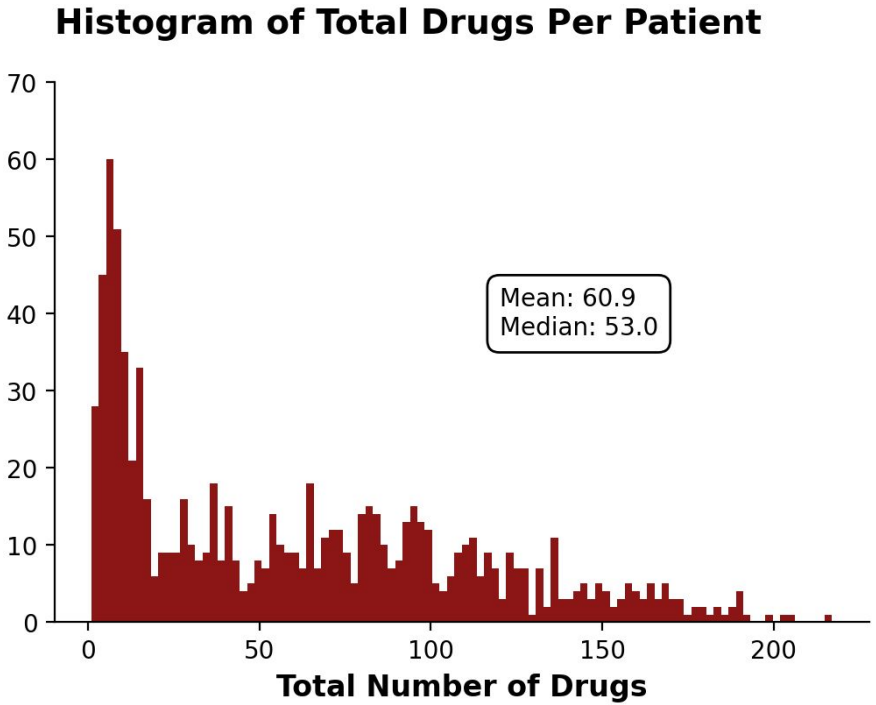
- Florida had the highest incidence of Congestive Heart Failure (concept ID 319835) across all years with 157 occurrences.
- I assume “incidence of Congestive Heart Failure” is separate from incidence of conditions co-occurring with Congestive Heart Failure (e.g., Hypertensive heart disease with Congestive Heart Failure).

5a. What are the 5 most commonly used drugs?

- The five most commonly used drugs are as follows:
  - Epoetin Alfa (1059)
  - Influenza virus vaccine, trivalent (IIV3), split virus, 0.5 mL dosage, for intramuscular use (619)
  - Paricalcitol Injectable Solution (553)
  - Oxygen 99 % Gas for Inhalation (550)
  - Omeprazole 20 MG Delayed Release Oral Capsule (345)
- The most common concept id in the drug exposures table was associated with “No matching concept.” I excluded it for the purposes of this analysis.
- I assume that the same drug at different doses can be considered a different drug—this way we do not lose any information. I interpret the "most commonly used drugs" as those with the highest number of entries in the drug exposure table.

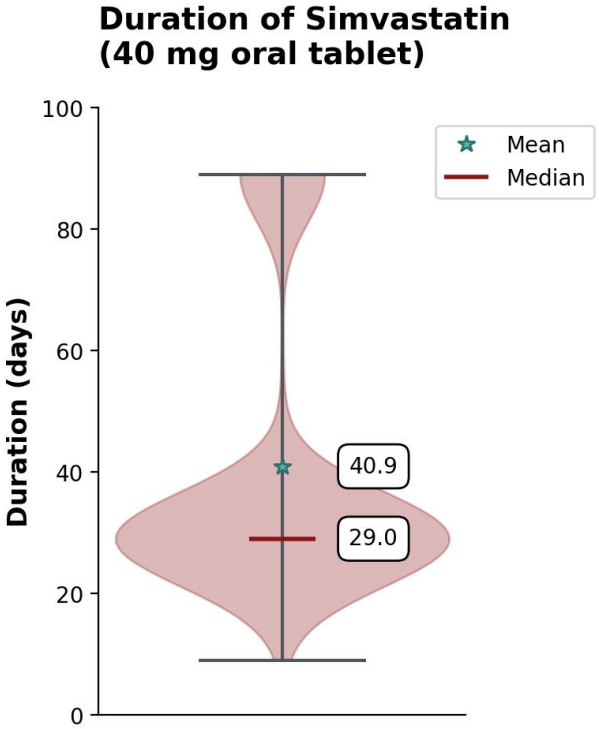
5b. What is the average number of drugs per person?

- The average number of (unique) drugs per person is 60.9 with a median of 53.0.
- This distribution seems to be influenced by some outlying patients who take many drugs.



5c. What is the average duration that the Simvastatin 40 mg oral tablet is taken for?

- The mean duration that the Simvastatin 40 mg oral tablet is taken is 40.9 days with a median of 29.0 days.
- I included data from both “Simvastatin 40 mg oral tablet (Zocor)” and “Simvastatin 40 mg oral tablet” in the above calculation.



Appendix

- The prompt stated that this dataset would be used for a hypothetical machine learning model. If collaborating on this, I might use engineered features (e.g., number of drugs taken or number of conditions) with the demographic variables in the person table as predictors of disease occurrence or mortality—though there may be too little death data.
- If there were clinical notes, I would employ traditional NLP or LLMs (if local servers are available) to extract features from the text. I might also explore integrating findings from omics and biometric data into the model.