HAP720-001: Data Integration Project - Phase 3

Name: Neha Rao
Gnumber: G01477370

1) Compare the performance (e.g., number of matched or mismatched) between MPI and patient matching

A) The following results can be used to compare the performance of patient matching and MPI (Master Patient Index):

Patient Matching Results:

- Matched Records: 1,110,886
- Mismatched Records: 3,079,264

Comparison of MPI: If MPI were accessible, it would probably offer:

- Increased Match Rates: An MPI keeps track of each patient's unique identifier across systems, making matches more accurate and dependable even when there are slight variations in clinical or demographic information.
- Reduce Mismatched Records: MPI can prevent issues like formatting variances and data entry errors that arise when matching is based on clinical and demographic variables.

2) Explain why the matching is not perfect?

A) The matching is not perfect due to several challenges inherent in the data and the methodology used. One primary issue is data inconsistencies, such as variations in formatting across datasets—for example, differences in how gender is recorded (e.g., "M" versus "Male") or variations in the representation of ICD codes. Data quality issues, including missing values and typographical errors, further complicate the process, as incomplete or incorrect information can prevent a match. Additionally, real-world variability, such as changes in patient information over time (e.g., age or address), can lead to discrepancies between records. The reliance on exact matching criteria, such as age, gender, and ICD codes, limits flexibility, as even minor differences result in mismatches. Finally, some unmatched records may genuinely reflect individuals not present in both datasets, especially if the sources differ in scope or timeframes. These factors collectively highlight the need for standardized data practices and more advanced matching techniques to improve accuracy.

Flexibility is restricted by the use of precise matching criteria, such as age, gender, and ICD codes, since even little variations lead to inconsistencies. Lastly, if the sources have different scopes or time periods, some unmatched records might really represent people who aren't in both datasets. All of these elements point to the necessity of more sophisticated matching strategies and standardized data processes in order to increase accuracy.