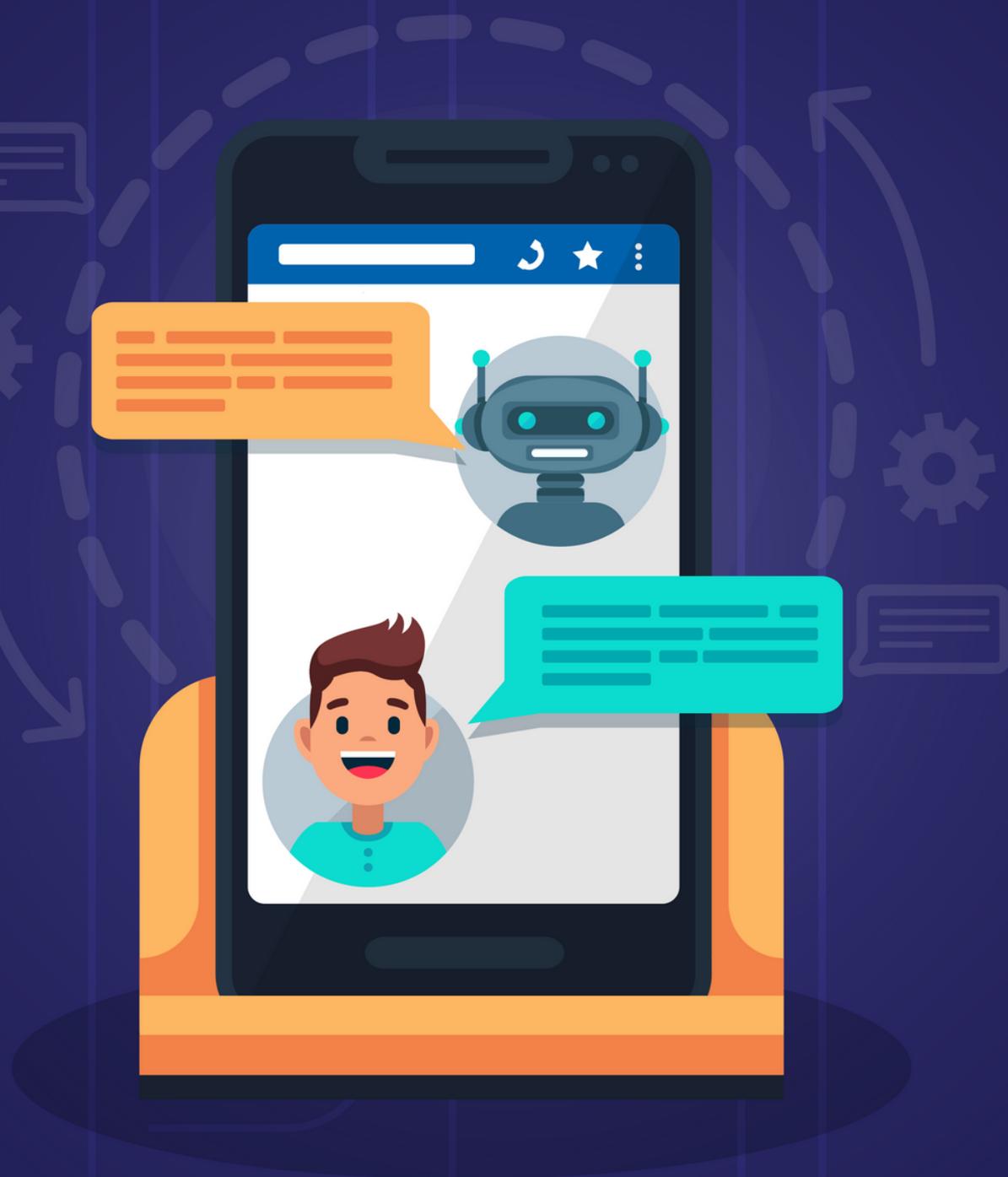


# Exploring, Designing & Generating SQL queries for Cohort Analysis based on CMS Healthcare Data



# Step: 1

Problem Statement : We need to create a cohort analysis tool to generate insights by exploring and designing the tool and finally by generating SQL queries for further analysis of the cohorts.

## 1. Steps to build a cohort using Claims Data and Electronic Health Records

AN So how can i define a cohort based on Claims Data and Electronic Health Record

To define a cohort based on Claims Data and Electronic Health Record (EHR) data, you would typically follow these steps:

1. Identify the population of interest: Define the population of interest based on the research question, such as individuals with a specific diagnosis or individuals who have received a specific treatment.
2. Extract the relevant data: Use the claims data to identify individuals who meet the inclusion criteria for the cohort, such as having a specific diagnosis code or procedure code. Extract the relevant data from the EHR, such as patient demographics, comorbidities, and laboratory results.
3. Clean and merge the data: Clean the data by removing any duplicate records, correcting errors, and ensuring that the data is consistent across both datasets. Merge the claims data and EHR data into a single dataset, so that all the relevant information for each individual is in one place.
4. Define the cohort: Using the merged data, define the cohort by applying the inclusion and exclusion criteria. For example, you might include individuals who have a specific diagnosis code and have received a specific treatment, and exclude individuals who have certain comorbidities or have received certain other treatments.
5. Validate the cohort: Validate the cohort by comparing it to external data sources, such as administrative data or other electronic health records, to ensure that the cohort is representative of the population of interest.

**Implementation Review :** Chat GPT is able to specifically define cohorts based on mentioned criteria and clearly identifies the process involved in defining the cohort.



# Step: 2

## 2. Identifying the process of Building a cohort around the Mortality of patients using Claims Data and Quality Data (Specific Requirement)



The fourth step in building a cohort around the mortality of patients using Claims Data and Quality Data is to define the cohort by applying the inclusion and exclusion criteria. This involves identifying the specific characteristics that are required for an individual to be included in the cohort and those that would exclude an individual from the cohort.

1. Inclusion criteria: These are the characteristics that an individual must have in order to be included in the cohort. For example, you might include individuals who have a specific diagnosis code, such as heart failure, and who have received a specific treatment, such as a heart transplant.
2. Exclusion criteria: These are the characteristics that would exclude an individual from the cohort. For example, you might exclude individuals who have certain comorbidities, such as cancer, or who have received certain other treatments, such as a different heart procedure.
3. Sub-cohort: After defining the inclusion and exclusion criteria, you can divide the cohort into two sub-cohorts: one sub-cohort with the individuals who are deceased, and another with those who are still alive.
4. Characteristics of sub-cohort: Compare the characteristics of the deceased and alive sub-cohort, and look for differences in treatment, comorbidities, and other factors that may be associated with mortality. This can include things like age, sex, race, and socioeconomic status, as well as details about the treatment received and any complications that may have occurred.
5. Quality checks: Perform quality checks to make sure that the data is accurate, complete, and unbiased. This can include checking for missing data, outliers, and errors. It's important to ensure that the cohort is representative of the population of interest and that the data is of high quality in order to make valid conclusions.



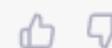
**Implementation Review :** For a specific requirement of defining cohort around Mortality, ChatGPT is able to mention all the steps required such as Inclusion Criteria, Exclusion Criteria, Sub Cohorts, Quality Checks etc. It was not able to provide steps relevant to CMS Healthcare data by providing which criteria to include/exclude.

# Step: 3

## 3. Designing a Cohort Analysis Tool based on CMS Healthcare Data



When using CMS healthcare data, a cohort analysis tool can generate several specific insights, including:



1. **Treatment effectiveness:** By analyzing utilization rates of specific treatments or procedures across different cohort groups, the tool can help to evaluate the effectiveness of different treatments and inform clinical decision making.
2. **Patient outcomes:** By tracking patient outcomes, such as readmission rates and length of stay, the tool can provide insights into the quality of care and inform efforts to improve patient outcomes.
3. **Provider performance:** By comparing utilization and patient outcome data across different providers, the tool can provide insights into provider performance and inform efforts to improve quality of care.
4. **Resource utilization:** By tracking utilization of specific healthcare services and resources, such as imaging or specialist visits, the tool can inform efforts to optimize resource utilization and reduce waste.
5. **Cost analysis:** By comparing the cost of care across different cohort groups and providers, the tool can support cost analysis and inform efforts to reduce healthcare spending.
6. **Population health management:** By tracking utilization and patient outcomes data for specific populations, such as elderly or chronic disease patients, the tool can support population health management efforts and inform efforts to improve the health of specific populations.



**Implementation Review :** ChatGPT is able to generate specific insights needed for designing the Cohort Analysis Tool. It was able to provide a broader structure of all the insights that can be included while designing a tool.

# Step: 4

## 4. Designing Cohort Analysis Tool (Specific Requirement)

AN

How to create this tool based on Treatment Effectiveness



To create a cohort analysis tool based on treatment effectiveness using CMS healthcare data, you can follow these steps:  

1. Define the cohort: Start by defining the criteria for creating the cohort groups, such as patients who received a specific treatment or procedure, patients with a specific diagnosis, or patients of a specific age group.
2. Collect the data: Next, collect the necessary data, including patient demographics, treatment information, utilization data, and patient outcomes data. Make sure the data is accurate and complete.
3. Clean and format the data: Clean the data to ensure it is accurate and complete, and format the data into a usable format for the analysis, such as a spreadsheet or database.
4. Analyze the data: Use the cohort analysis tool to perform the analysis, focusing on metrics related to treatment effectiveness, such as utilization rates, readmission rates, length of stay, and patient outcomes.
5. Visualize the results: Use charts, tables, or graphs to visualize the results, making sure the data is presented in a clear and meaningful way.
6. Interpret the results: Analyze the results to draw conclusions about the treatment effectiveness, such as the most effective treatments for specific diagnoses or patient populations, and the factors that contribute to the treatment's effectiveness.
7. Refine and repeat: Repeat the process, refining the criteria for creating the cohort groups and adjusting the analysis as needed, until you have a complete and accurate picture of the treatment effectiveness.



**Implementation Review :** For designing Cohort Analysis Tool for a specific requirement, ChatGPT was clearly able to define and explain the steps required for creating the tool. However, these steps provided only a general idea of creating a tool and was unable to provide implementation steps which can directly be used for creating the tool.



# Step: 5.1

## 5. Basic Cohort Analysis Via ChatGPT on OMOP data Model

AN I want to generate cohort analysis based on omop data model. Suggest some sql queries to perform this operation

 Sure, here are a few SQL queries that can help you perform a basic cohort analysis using the OMOP data model:

1. Extracting the demographics of the cohort:

sql

 Copy code

```
SELECT person_id, gender_concept_id, year_of_birth, race_concept_id, ethnicity_concept_id
FROM person
```

1. Counting the number of individuals in the cohort:

sql

 Copy code

```
SELECT COUNT(DISTINCT person_id)
FROM person
```

1. Counting the number of unique conditions for each individual in the cohort:

vbnet

 Copy code

```
SELECT person_id, COUNT(DISTINCT condition_occurrence.condition_concept_id)
FROM condition_occurrence
JOIN person ON condition_occurrence.person_id = person.person_id
```

 Regenerate response

# Step: 5.2

```
--Demographics of Cohort
SELECT top 3 person_id, gender_concept_id, year_of_birth, race_concept_id, ethnicity_concept_id
FROM [dbo].[person1]

--Number of individuals in cohort
SELECT COUNT(DISTINCT person_id) as Total_Patients
FROM [dbo].[person1]

--Counting the number of unique conditions for each individual in cohort
SELECT a.[PERSON_ID], COUNT(DISTINCT [CONDITION_OCCURRENCE_ID]) as Unique_Conditions
FROM [dbo].[Cond Occurrence1]
JOIN [dbo].[person1] a ON [dbo].[Cond Occurrence1].[PERSON_ID] = a.[PERSON_ID]
GROUP BY a.[PERSON_ID]
```

90 %

Results Messages

	person_id	gender_concept_id	year_of_birth	race_concept_id	ethnicity_concept_id
1	6	8532	1963	8516	0
2	123	8507	1950	8527	0
3	129	8507	1974	8527	0

	Total_Patients
1	2694

	PERSON_ID	Unique_Conditions
1	1	2
2	1000	2
3	101	2
4	1017	2
5	1018	2



**Implementation Review :** We are able to successfully execute the queries apart from a minor ambiguity error. It provides only the basic query on which we have to modify the naming conventions which we want as end result.

# Step: 6.1

6. 1. Asking ChatGPT to provide inclusion and exclusion criteria to specify a cohort according to constraints.



Sure, here's the equivalent query in MS SQL Server syntax for the inclusion criteria example:  

sql

 Copy code

```
SELECT person.person_id, person.gender_concept_id, person.year_of_birth, person.race
FROM person
WHERE year_of_birth BETWEEN (YEAR(GETDATE()) - 80) AND (YEAR(GETDATE()) - 18)
AND person.person_id IN (SELECT condition_occurrence.person_id FROM condition_occurrence)
```

And here's the equivalent query in MS SQL Server syntax for the exclusion criteria example:

sql

 Copy code

```
SELECT person.person_id, person.gender_concept_id, person.year_of_birth, person.race
FROM person
WHERE year_of_birth BETWEEN (YEAR(GETDATE()) - 80) AND (YEAR(GETDATE()) - 18)
AND person.person_id IN (SELECT condition_occurrence.person_id FROM condition_occurrence)
AND person.person_id NOT IN (SELECT drug_exposure.person_id FROM drug_exposure WHERE drug_exposure.person_id IS NOT NULL)
AND person.person_id NOT IN (SELECT observation.person_id FROM observation WHERE observation.person_id IS NOT NULL)
```

# Step: 6.2

```
-- Inclusion Criteria : Age (Patients between 18 and 80 years age) and Diagnosis (Diagnosed with specific condition )
SELECT a.person_id, a.gender_concept_id, a.year_of_birth, a.race_concept_id, a.ethnicity_concept_id
FROM [dbo].[person1] a
WHERE year_of_birth BETWEEN (YEAR(GETDATE()) - 80) AND (YEAR(GETDATE()) - 18)
AND a.person_id IN (SELECT b.person_id FROM [dbo].[Cond Occurrence1] b WHERE b.condition_concept_id = 40481087)

--Exclusion Criteria : Patients who have received a specific treatment shoul not be included
SELECT a.person_id, a.gender_concept_id, a.year_of_birth, a.race_concept_id, a.ethnicity_concept_id
FROM [dbo].[person1] a
WHERE year_of_birth BETWEEN (YEAR(GETDATE()) - 80) AND (YEAR(GETDATE()) - 18)
AND a.person_id IN (SELECT b.person_id FROM [dbo].[Cond Occurrence1] b WHERE b.condition_concept_id = 40481087)
AND a.person_id NOT IN (SELECT person_id FROM [dbo].[drug exposure1] WHERE drug_concept_id = 920300)
```

90 %

Results Messages

	person_id	gender_concept_id	year_of_birth	race_concept_id	ethnicity_concept_id
1	18	8532	1965	8527	0
2	114	8532	1972	8527	0
3	35	8532	1960	8527	0
4	40	8507	1951	8527	0
5	90	8532	1961	8527	0
6	135	8532	1971	8527	0

	person_id	gender_concept_id	year_of_birth	race_concept_id	ethnicity_concept_id
1	40	8507	1951	8527	0
2	90	8532	1961	8527	0
3	135	8532	1971	8527	0
4	61	8532	1968	8527	0
5	66	8532	1958	8527	0



**Implementation Review:** ChatGPT is able to effectively develop SQL logic for creating both exclusion and inclusion logic. We can easily create specific cohorts as per the demand. However, it does not provide an efficient way of referring the tables in queries which are needed to be modified while implementing the query in SQL.

# ChatGPT efficiency in SQL:



## Efficient

- Clearly Defining Steps and Explaining those steps
- Generating a clear process and providing all the parameters necessary for a specific requirement
- Generates valid and accurate queries as per the domain specified.
- Basic and Advanced SQL queries are generated as per the modifications in the requirements which are ready to implement.



## Partially Efficient

- Not able to provide all the steps included for a specific requirement
- Finds difficult to provide any output which is relevant to any specific database.
- General Steps are generated instead of Implementation steps for qualitative data.

# Authors



**Ani Kannal**  
**Head of Data Science and Analytics**  
**Healthark Insights**



**Ankush Jain**  
**Analytics Consultant**  
**Healthark Insights**

**CONTACT US**

821, Sun Avenue One  
Manekbag, Ambawadi,  
Amdedabad, India  
contact@healtharkinsights.com  
healtharkinsights.com