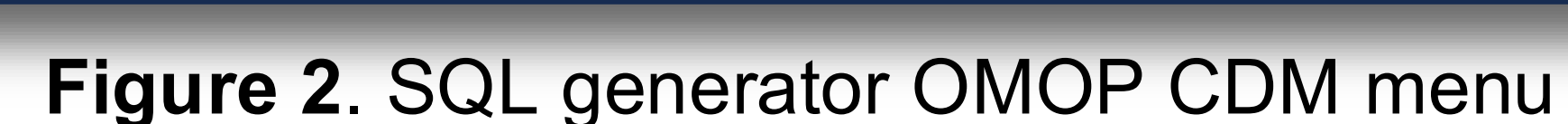


## Background:

Querying OMOP databases remains a practical bottleneck: many clinically meaningful questions require non-trivial SQL (multiple joins, temporal logic, inclusion/exclusion criteria, and concept-set logic), plus familiarity with OMOP conventions. In parallel, platforms such as OpenAI GPTs enable the creation of custom conversational agents that combine instructions, structured guidance, and domain knowledge, which is particularly relevant for specialised schemas like OMOP CDM. However, in clinical research, using free-form text-to-SQL can lead to mistakes. To address this gap, we developed a publicly accessible custom GPT that supports users in generating PostgreSQL queries over OMOP CDM using natural language plus a guided, menu-based interaction pattern. This approach aims to democratise access to OMOP data for feasibility analyses and cohort extraction while ensuring that outputs remain aligned with OMOP structure and conventions in federated and multi-institutional research contexts, such as the European Health Data Space platforms.

## Results

Figure 1 shows the custom ChatGPT created. A dedicated web interface for the “SQL generator OMOP-CDM” custom GPT was deployed, providing one-click example prompts and a menu-driven workflow to help users generate OMOP-CDM PostgreSQL queries. Figure 2 displays the first menu of the SQL generator.



## Methods

- Assistant Development:** A conversational agent was developed based on OpenAI's custom GPT 5.4 technology and made publicly accessible. The system helps users define groups and criteria using structured menus and automatically converts a plain-language question into an SQL statement that matches the OMOP-CDM structure. It is applied to a complete, standardised OMOP CDM schema in PostgreSQL:

- The "OMOP-CDM Explorer" assistant helps users explore the OMOP CDM database by using a menu to perform tasks such as checking demographics, analysing conditions and drugs, defining groups, looking at trends over time, and searching for vocabulary.
- To improve the accuracy and portability of natural-language-to-SQL generation, we embedded an explicit representation of the OMOP schema in the assistant's knowledge base:
  - `omop_schema.sql`: the source of truth (tables, columns, datatypes, PK/FK constraints)
  - `omop_schema_index.md`: human-readable map of the schema
  - `omop_schema_index.js`: machine-readable index

**2 Evaluation:** We performed several tests using multiple queries to gauge the assistant's feasibility and usefulness. We designed a set of natural language cohort requests reflecting use cases a clinical researcher might have and queried our database. Below an example of the table generated based on the query created by the SQL generator (using the generated query of 135 lines):

|    | section                  | label  | concept_id | disease_name                                  | person_count | occurrence_count |
|----|--------------------------|--------|------------|-----------------------------------------------|--------------|------------------|
| 1  | gender_breakdown         | FEMALE | (NULL)     | (NULL)                                        | 865,433      | (NULL)           |
| 2  | gender_breakdown         | MALE   | (NULL)     | (NULL)                                        | 926,209      | (NULL)           |
| 3  | top_conditions_by_gender | FEMALE | 320,128    | Essential hypertension                        | 69,897       | 188,837          |
| 4  | top_conditions_by_gender | FEMALE | 81,902     | Urinary tract infectious disease              | 45,992       | 69,429           |
| 5  | top_conditions_by_gender | FEMALE | 194,133    | Low back pain                                 | 30,927       | 46,338           |
| 6  | top_conditions_by_gender | FEMALE | 375,545    | Cataract                                      | 30,503       | 77,905           |
| 7  | top_conditions_by_gender | FEMALE | 4,014,295  | Single live birth                             | 29,347       | 39,107           |
| 8  | top_conditions_by_gender | MALE   | 320,128    | Essential hypertension                        | 61,596       | 170,011          |
| 9  | top_conditions_by_gender | MALE   | 433,071    | Conjuction of multiple sites                  | 30,504       | 35,181           |
| 10 | top_conditions_by_gender | MALE   | 4,193,704  | Type 2 diabetes mellitus without complication | 30,272       | 92,801           |
| 11 | top_conditions_by_gender | MALE   | 194,133    | Low back pain                                 | 29,087       | 42,204           |
| 12 | top_conditions_by_gender | MALE   | 437,264    | Tobacco dependence syndrome                   | 28,011       | 51,182           |



Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute  
and Data Science Unit, Hospital del Mar Barcelona (Spain)

Angela Leis, Miguel-Angel Mayer, Juan Manuel Ramírez-Anguita, Carla Arxé

EUROPEAN OHDSI Symposium 18-20 April 2026  
Rotterdam

