Machine Learning-assisted query and information retrieval system on real-world data

PRESENTER: **Miguel Romero Calvo**
Yupeng Li

Introduction

How many people were diagnosed with Alzheimer's disease

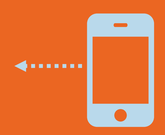
Real-world data

Technical barriers
- Database
- Data structure
- Customized tools
- Programming

19,570

Directly query OMOP-CDM datasets with natural language

- making the data querying as easy as a google search

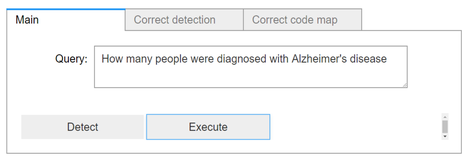


Take a picture or click the link to learn more details

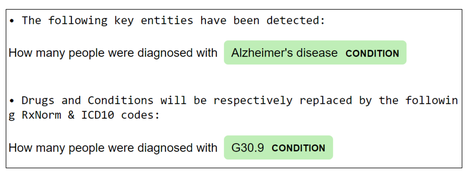
https://github.com/OHDSI/NOTOS

Workflow

1. User Input

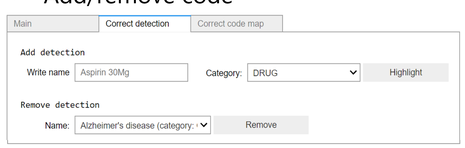


2. Entity recognition by Amazon Comprehend Medical (CM)

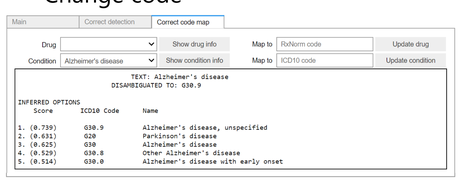


3. Code correction (optional)

- Add/remove code



- Change code



4. Text pre-processing

```
How many people were diagnosed with
<ARG-CONDITION><0>
```

5. Text-to-SQL translation

```
SELECT COUNT( DISTINCT con1.person_id
) FROM (<SCHEMA>.condition_occurrence
con1 JOIN <CONDITION-TEMPLATE><ARG-
CONDITION><0> ON
con1.condition_concept_id=concept_id);
```

6. SQL post-processing

```
SELECT COUNT( DISTINCT con1.person_id ) FROM
(omop_synpuf_110k.condition_occurrence con1 JOIN (
SELECT descendant_concept_id AS concept_id FROM
(SELECT * FROM (SELECT concept_id_2 FROM ( (SELECT
concept_id FROM  omop_synpuf_110k.concept WHERE
vocabulary_id='ICD10CM' AND ( concept_code='G30.9' ))
JOIN  ( SELECT concept_id_1, concept_id_2 FROM
omop_synpuf_110k.concept_relationship WHERE
relationship_id='Maps to' )  ON
concept_id=concept_id_1) ) JOIN
omop_synpuf_110k.concept ON concept_id_2=concept_id)
JOIN omop_synpuf_110k.concept_ancestor ON
concept_id=ancestor_concept_id )  ON
con1.condition_concept_id=concept_id) ;
```

7. SQL execution

- Request run successfully ✅.

	count
0	19570

Text-to-SQL model

- Compiled a list of high-frequency questions and the corresponding SQL queries
- Generated multiple real-world variations for each pre-defined questions
- Fine-tuned the T5 model for text-to-SQL translation

Model accuracy for in-scope questions

Metric	Validation	Test
Exact Match	0.9926	0.9920
Execution	0.9999	0.9999

Miguel Romero Calvo, Yupeng Li, Tesfagabir Meharizghi, Weilin Meng, Selvan Senthivel, Saman Sarraf, Lin Lee Cheong

aws MERCK OHDSI