Name:	Chi Yuan
Affiliation:	Columbia University
Email:	cy2465@cumc.columbia.edu
Presentation type (s):	Software Demonstration

Criteria2Query: Automatically Transforming Clinical Research Eligibility Criteria Text to OMOP Common Data Model (CDM)-based Cohort Queries

Chi Yuan, MS[1,4], Patrick B. Ryan, PhD[1,2,3], Yixuan Guo, MA[1], Peng Jin, MS[1], Kang Tian, MA[1], Chunhua Weng, PhD[1]

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA; [2]Observational Health Data Sciences and Informatics, New York, NY, USA; [3]Janssen Research & Development, LLC, Titusville, NJ, USA; [4]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, P.R. China

Patient recruitment has been a persistent barrier to clinical and translational research[1]. Over 50% of studies fail due to difficulties in recruitment. The universal adoption of electronic health records makes it possible to query distributed, large clinical databases for prescreening potentially eligible patients for performing feasibility assessment or cohort identification for clinical studies. However, interpretations of clinical research eligibility criteria may vary from site to site or from person to person, which can lead to incompatible or mismatched patient queries and compromise the integrity of multi-site studies. In addition, encoding of eligibility criteria into database queries often involves laborious human effort, which is costly, not scalable, and often prohibitive for clinicians and researchers, who do not necessarily understand informatics and know how to map criteria concepts to their data representations. This study aims to build a natural language processing system to transform eligibility criteria text into standards-based cohort identification queries that are sharable and configurable by end users, who can then focus their energy on feasibility assessment and iterative refinement of the criteria based on dynamic feedback with patient counts during eligibility criteria design.

The system Criteria2Query consists of five modules (**Figure 1**): i.e., text parsing, criteria filtering, terminology standards-based concept mapping, automatic query formulation, and query execution with dynamic feedback generation for users. Criteria parsing is supported by an open-source machine learning parser, EliIE[2], which outputs XML representations for recognized entities and their temporal relationships using the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) V5[3]. Criteria (e.g., "*willing to sign the consent*" and "*able to walk 5 miles on treadmill*") that cannot be queried within EHR for prescreening purposes are filtered out based on heuristics or empirical knowledge. Concepts and their relationships in retained criteria are then mapped to the Observational Health Data Sciences and Informatics (OHDSI) controlled clinical vocabularies to obtain concept IDs for each concept and to create structured cohort definitions, which are further translated into cohort queries in JSON or SQL formats using a public OHDSI Web API (https://github.com/OHDSI/WebAPI). These queries can be executed against any OMOP CDM-based patient database to prescreen potentially eligible patients.

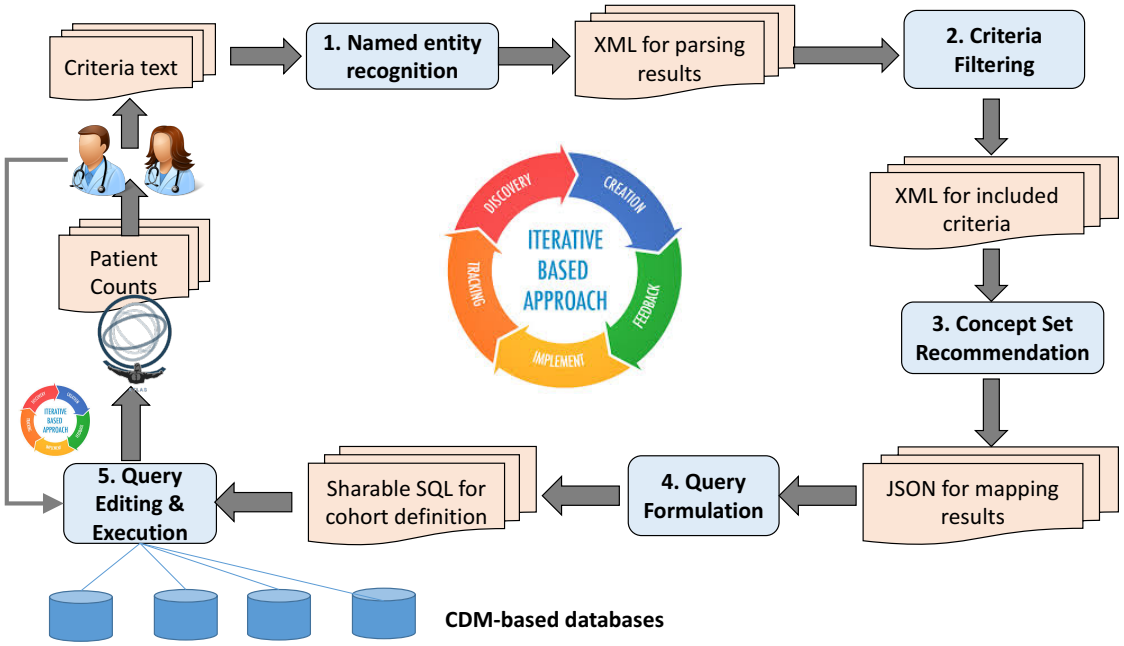

Figure 1. The behind-the-scene parsing pipeline that enables iterative eligibility criteria design.

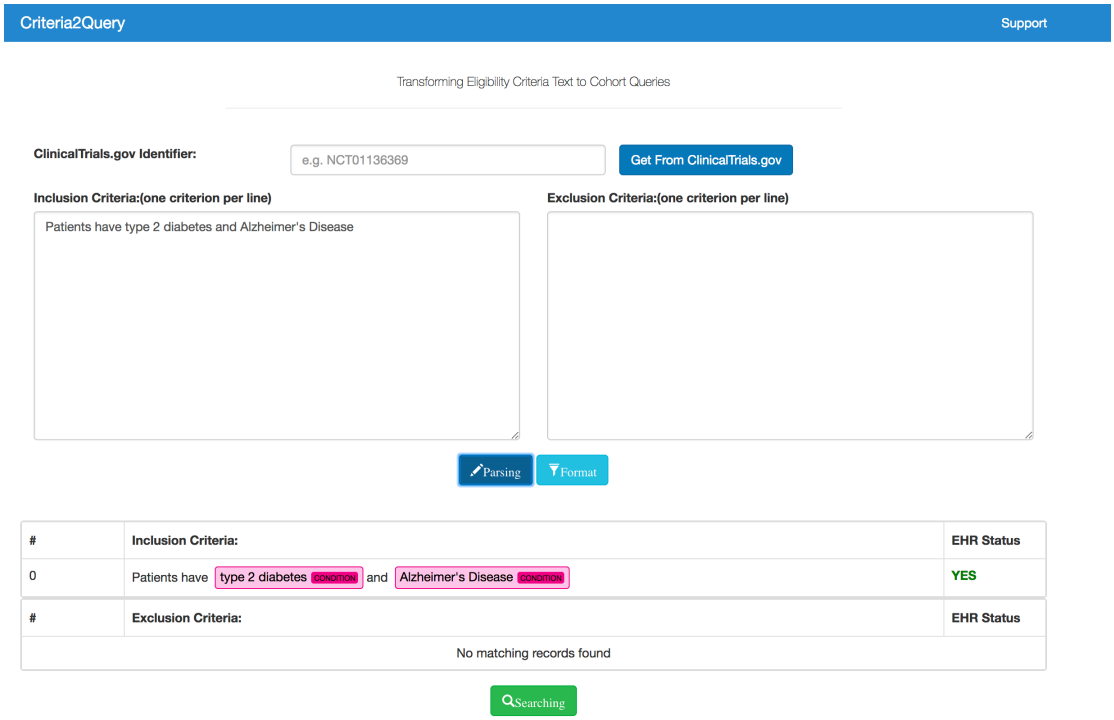

Figure 2. Snapshot of free-text parsing page.

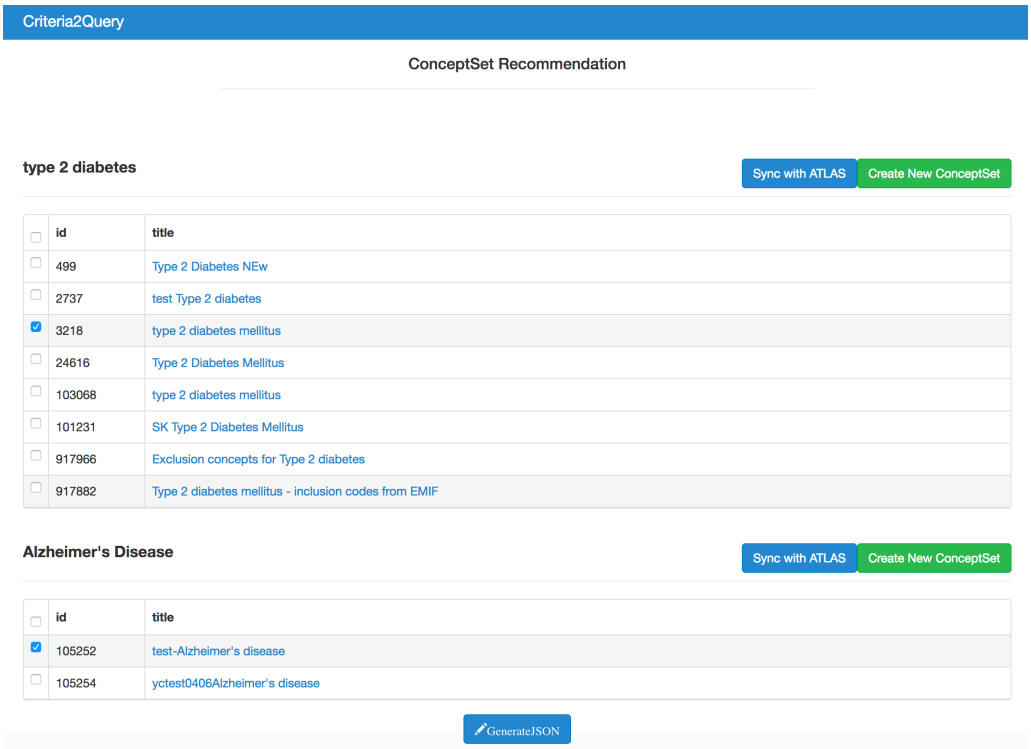

Figure 3. Snapshot of concept set mapping page.

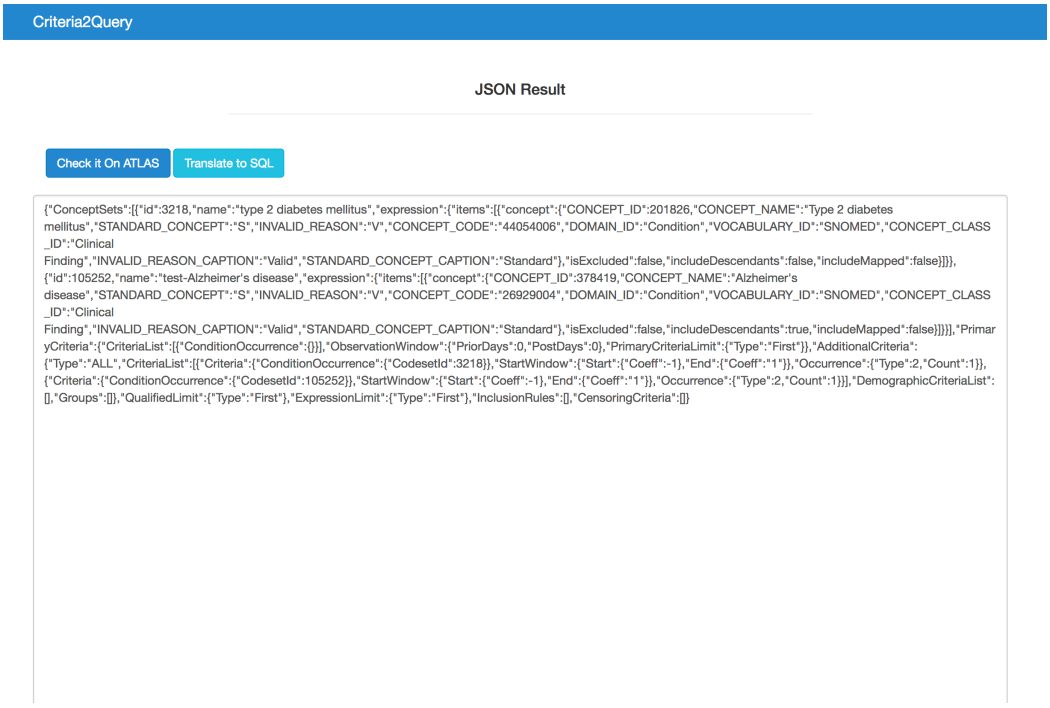

Figure 4. Snapshot of JSON result page.

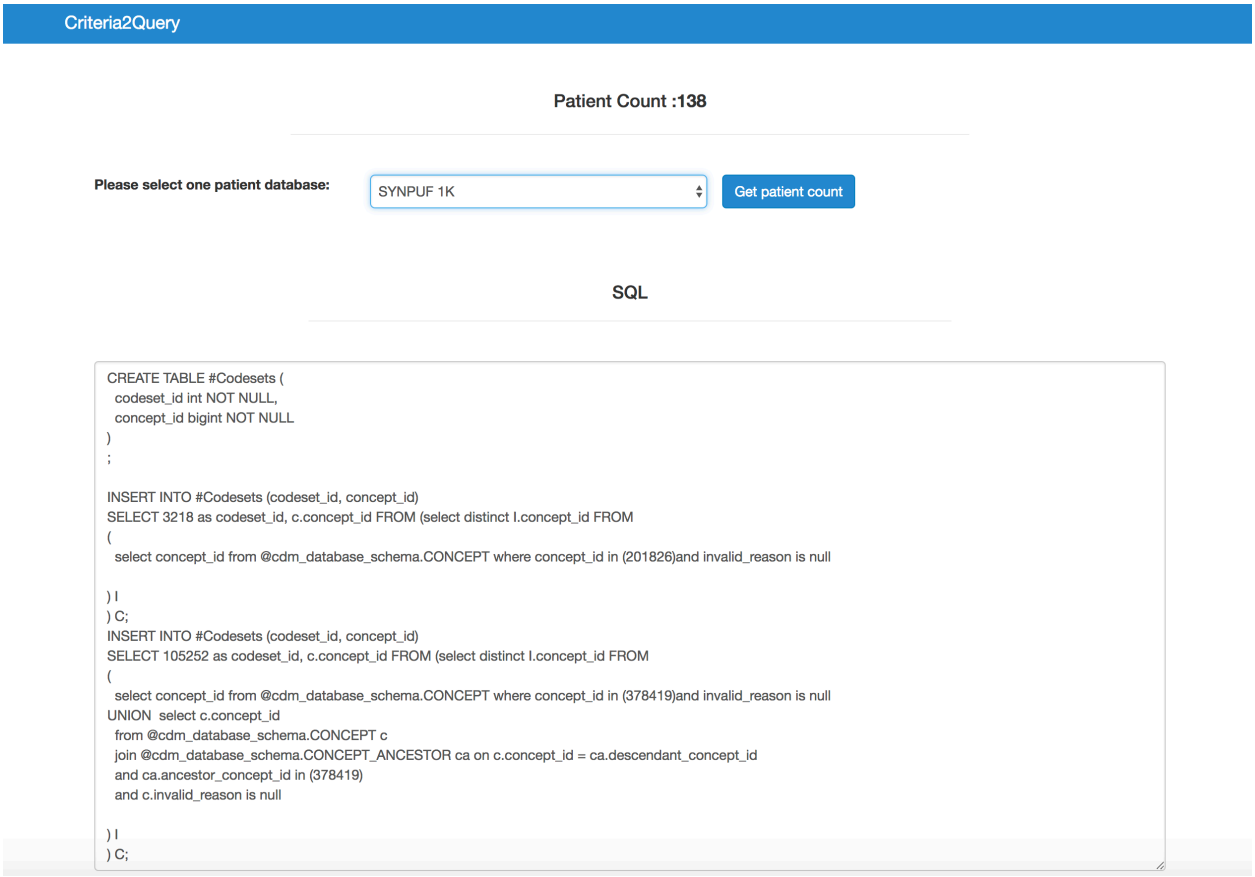

Figure 5. Snapshot of SQL and Patient count page.