OHDSI NLP WG Monthly Meeting

02/13/2019

Agenda

- Introduction of New Members
- Criteria2Query: a natural language interface to OMOP CDM databases for cohort identification – Chunhua Weng and Chi Yuan
- Ongoing projects
- Other issues

PRESENTATION

Criteria2Query: a natural language interface to OMOP CDM databases for cohort identification

Chunhua Weng and Chi Yuan

Criteria2Query: a natural language interface to OMOP CDM databases for cohort identification

Chi Yuan[1], Patrick B Ryan[1,2], Casey Ta[1], Yixuan Guo[1], Ziran Li[1], Jill Hardin[2], Rupa Makadia[2], Peng Jin[1], Ning Shang[1], Tian Kang[1], Chunhua Weng[1]

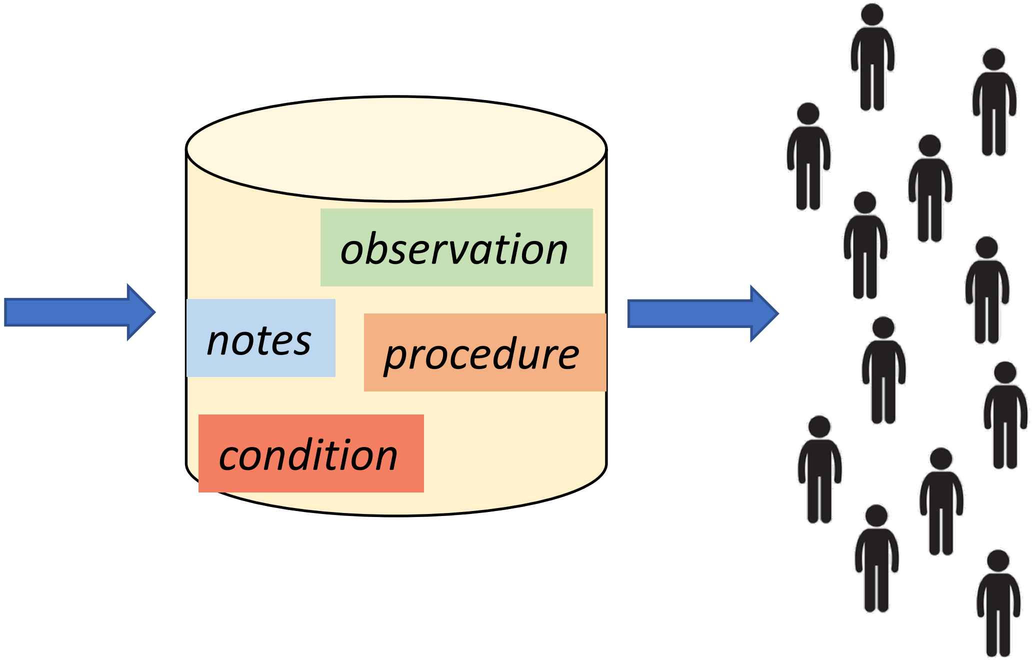
[1]Columbia University; [2]Janssen Inc.

February 12, 2019

Cohort identification

- Clinical diagnosis of ST-segment elevation acute myocardial infarction
- Must be treated within 12 hours after symptom onset
- Must be able to walk
- Must receive successful primary percutaneous coronary intervention

NCT01484158



Task breakdown

- Entity recognition: what is being searched for?
- Concept specification: what does it mean here?
- Concept mapping/normalization: how is it coded in a database?
- Phenotyping: what if the concept is implicitly represented?
- Data location: is it in the database? If yes, where? Which source is more reliable or convenient if there is > 1 source?
- Query formulation

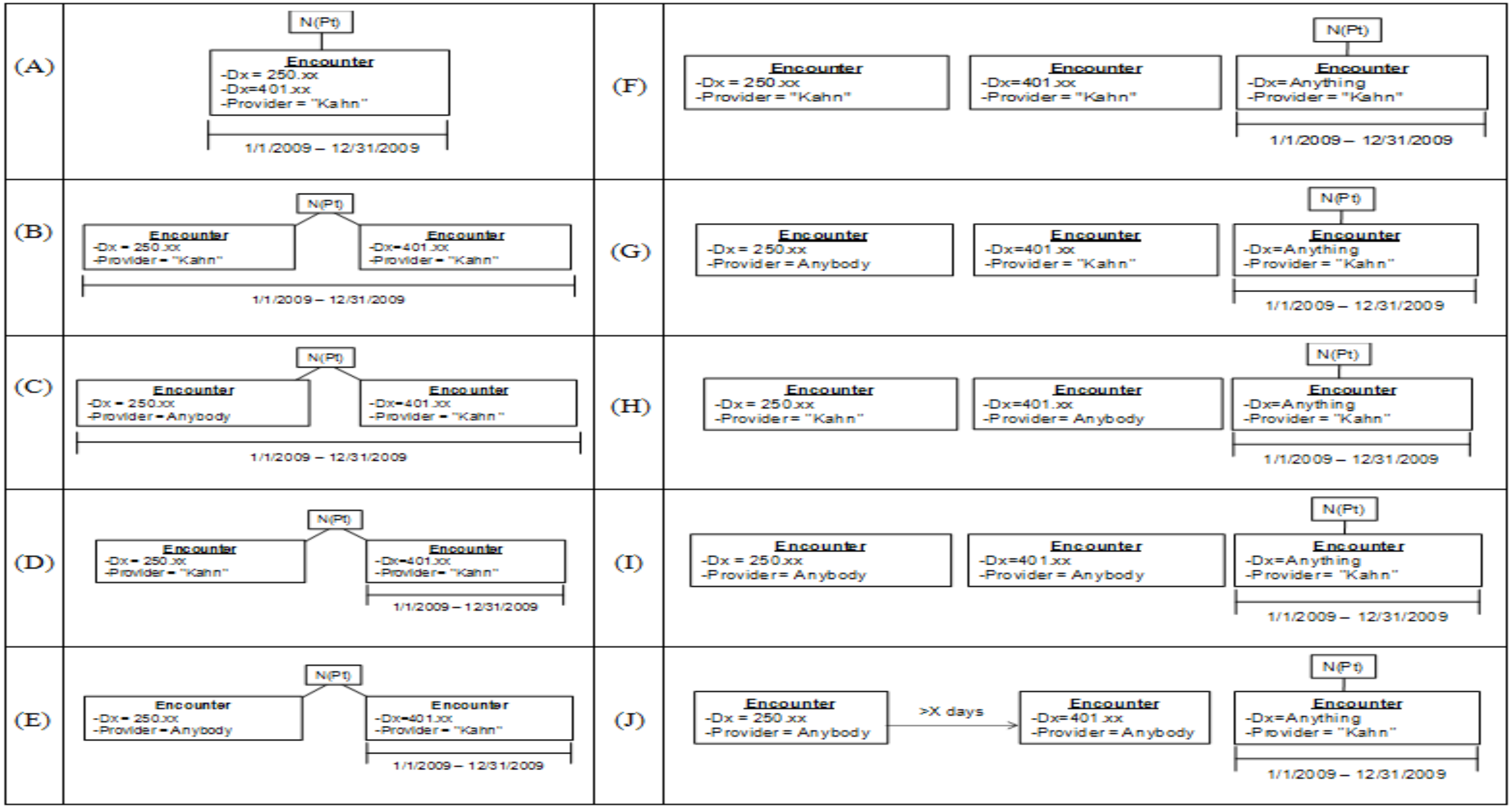
- Clinical diagnosis of ST-segment elevation acute myocardial infarction
- Must be treated within 12 hours after symptom onset
- Must be able to walk
- Must receive successful primary percutaneous coronary intervention

NCT01484158

Ten Translations for One Criterion

e.g., "*ambulatory patients seen by Dr. Michael Kahn with diabetes mellitus and essential hypertension between 1/1/2009 and 12/31/2009?*"

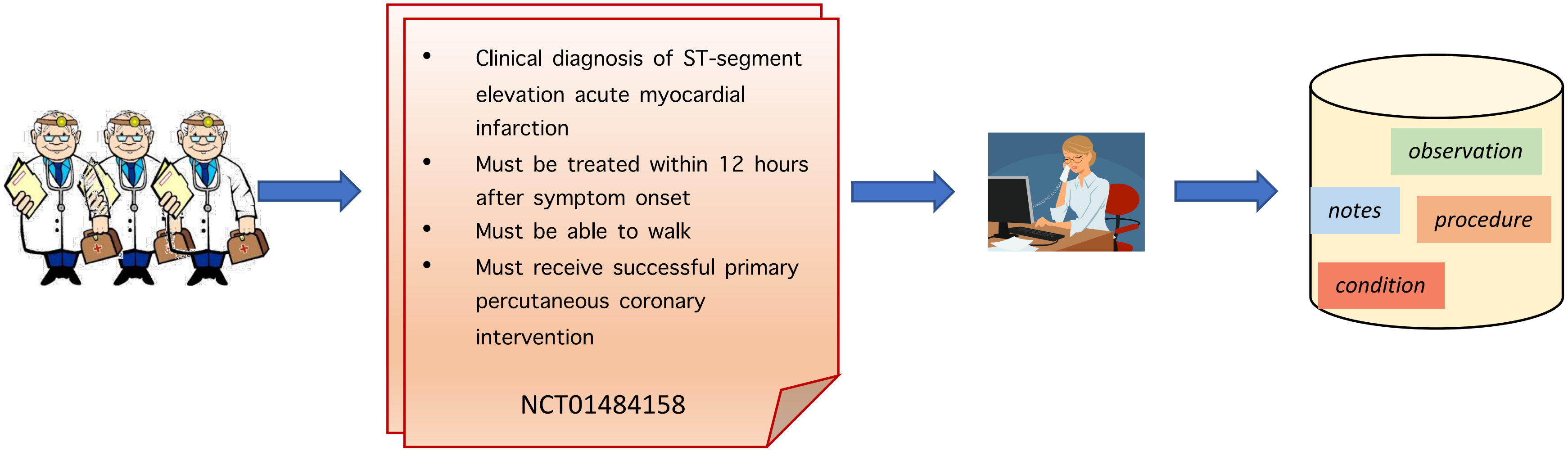
Table 1: Ten graphical diagrams representing the question: "How many ambulatory patients did I ("Provider = Kahn") see with diabetes mellitus (ICD-9 = 250.xx) and essential hypertension (ICD-9 = 401.xx) between January 1, 2009 and December 31, 2009?" Each diagram, when converted into a database query, returns a different result. N(Pt) = number of patients.



(A)	N(Pt) Encounter -Dx = 250.xx -Dx=401.xx -Provider = "Kahn" 1/1/2009 – 12/31/2009	(F)	Encounter -Dx = 250.xx -Provider = "Kahn" Encounter -Dx=401.xx -Provider = "Kahn" N(Pt) Encounter -Dx= Anything -Provider = "Kahn" 1/1/2009 – 12/31/2009
(B)	N(Pt) Encounter -Dx = 250.xx -Provider = "Kahn" Encounter -Dx=401.xx -Provider = "Kahn" 1/1/2009 – 12/31/2009	(G)	Encounter -Dx = 250.xx -Provider = Anybody Encounter -Dx=401.xx -Provider = "Kahn" N(Pt) Encounter -Dx=Anything -Provider = "Kahn" 1/1/2009 – 12/31/2009
(C)	N(Pt) Encounter -Dx = 250.xx -Provider = Anybody Encounter -Dx=401.xx -Provider = "Kahn" 1/1/2009 – 12/31/2009	(H)	Encounter -Dx = 250.xx -Provider= "Kahn" Encounter -Dx=401.xx -Provider= Anybody N(Pt) Encounter -Dx=Anything -Provider = "Kahn" 1/1/2009 – 12/31/2009
(D)	N(Pt) Encounter -Dx = 250.xx -Provider = "Kahn" Encounter -Dx=401.xx -Provider = "Kahn" 1/1/2009 – 12/31/2009	(I)	Encounter -Dx = 250.xx -Provider= Anybody Encounter -Dx=401.xx -Provider = Anybody N(Pt) Encounter -Dx=Anything -Provider= "Kahn" 1/1/2009 – 12/31/2009
(E)	N(Pt) Encounter -Dx = 250.xx -Provider = Anybody Encounter -Dx=401.xx -Provider = "Kahn" 1/1/2009 – 12/31/2009	(J)	Encounter -Dx = 250.xx -Provider = Anybody >X days Encounter -Dx=401.xx -Provider = Anybody N(Pt) Encounter -Dx=Anything -Provider = "Kahn" 1/1/2009 – 12/31/2009

--material from Dr. Michael G Kahn
Michael.Kahn@childrenscolorado.org

State of the art



- Clinical diagnosis of ST-segment elevation acute myocardial infarction
- Must be treated within 12 hours after symptom onset
- Must be able to walk
- Must receive successful primary percutaneous coronary intervention

NCT01484158

observation

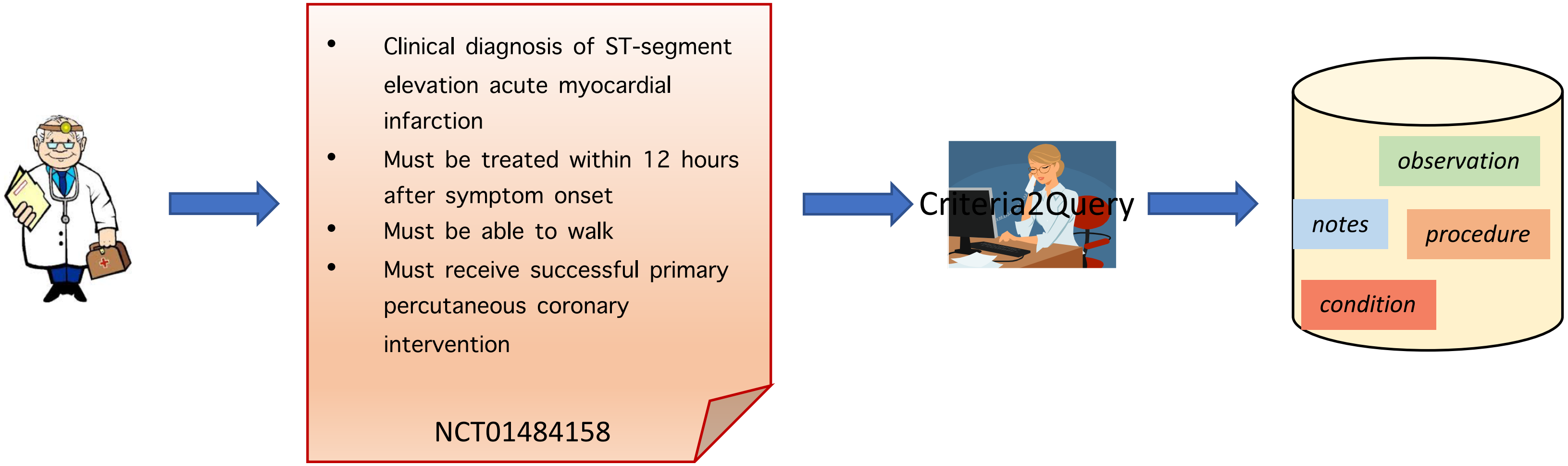
notes

procedure

condition

- High cost
- Long waiting time
- Fragmented knowledge
- Limited query reuse and knowledge sharing
- No autonomy for clinician scientist

The goal of Criteria2Query:
clinician autonomy with minimal effort



- Clinical diagnosis of ST-segment elevation acute myocardial infarction
- Must be treated within 12 hours after symptom onset
- Must be able to walk
- Must receive successful primary percutaneous coronary intervention

NCT01484158

Criteria2Query

observation

notes

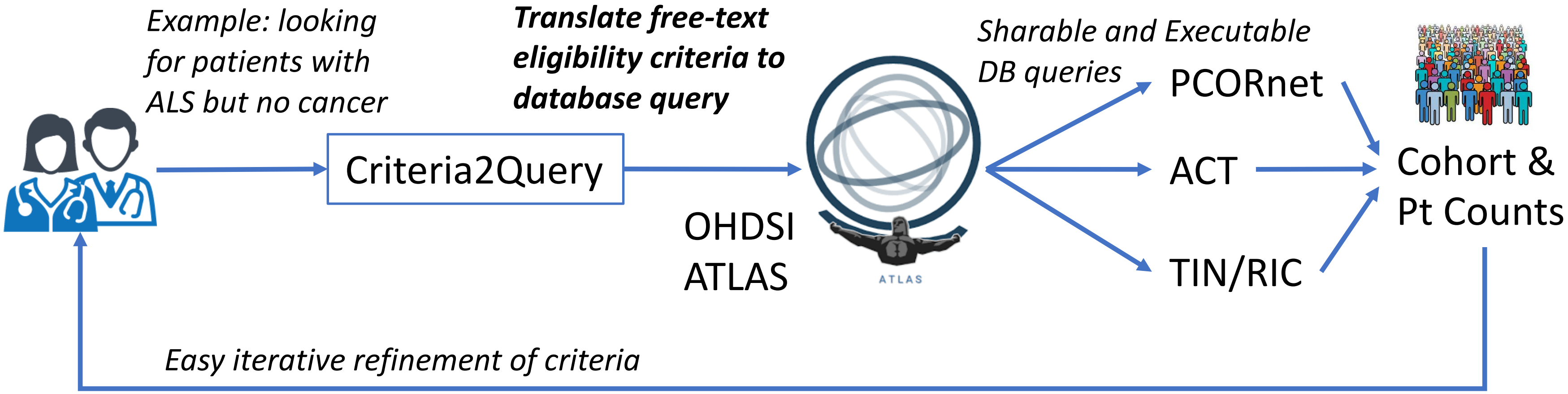
procedure

condition

Currently focus on information retrieval (anything queryable),
not on phenotype knowledge engineering (anything that needs knowledge or inference)

Criteria2Query for reusable and sharable queries

Goal: minimize time needed from clinicians to translate English concepts to codes in ICD-9, SNOMED, LOINC, RxNorm, and etc., used by databases and enables rapid iterative feasibility assessment



Example: looking for patients with ALS but no cancer

Criteria2Query

Translate free-text eligibility criteria to database query

OHDSI ATLAS

Sharable and Executable DB queries

PCORnet

ACT

TIN/RIC

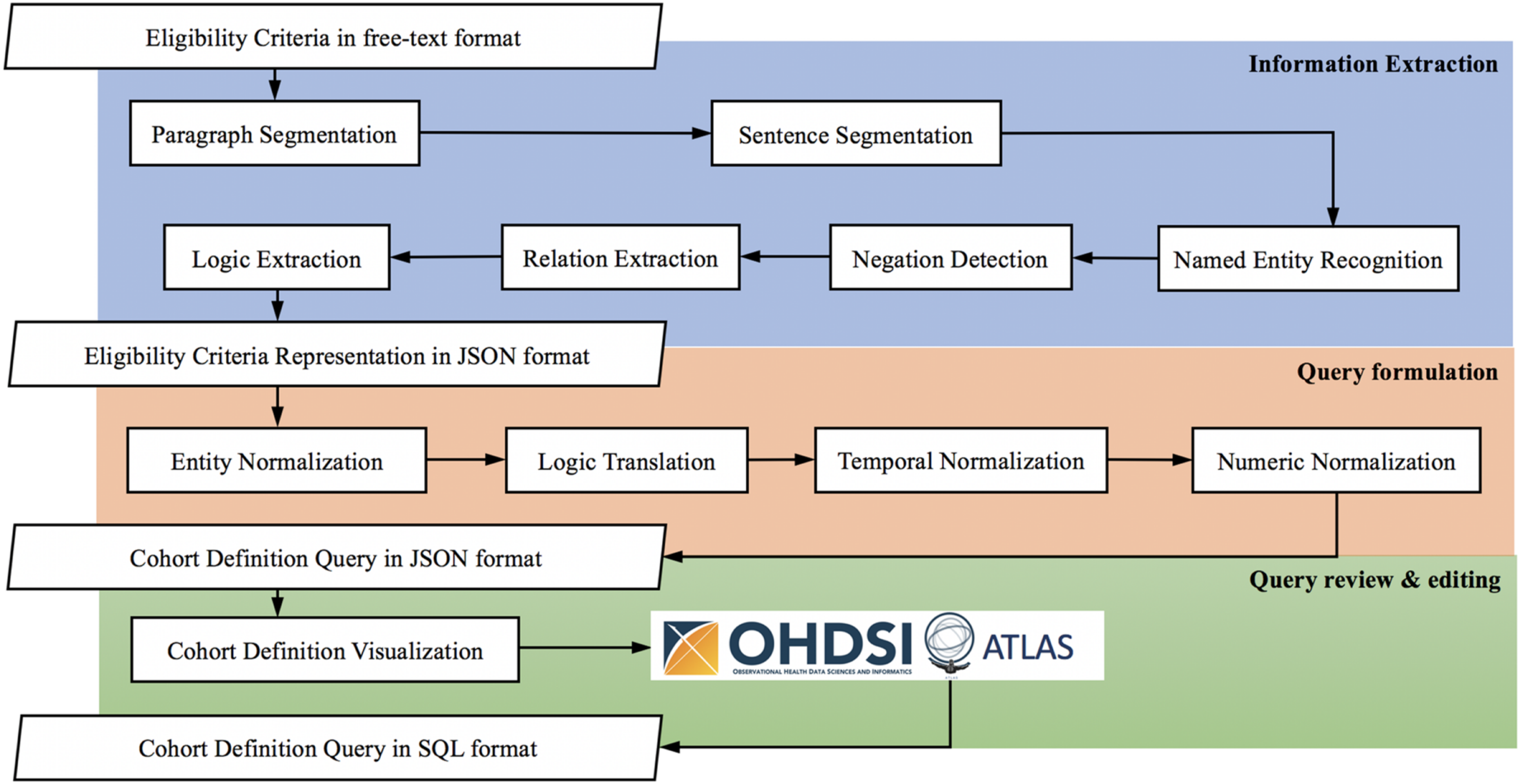
Cohort & Pt Counts

Easy iterative refinement of criteria

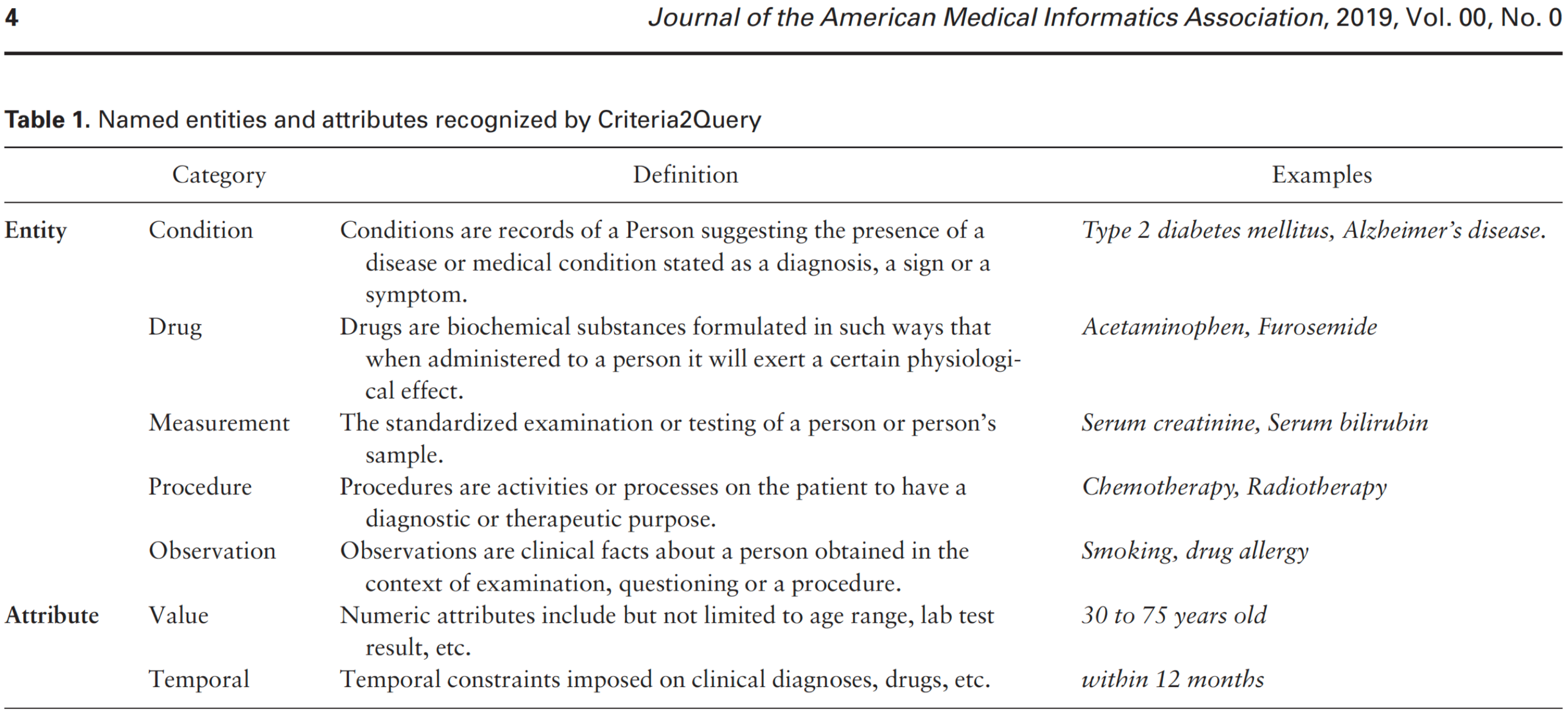
Brief demo

https://www.youtube.com/watch?v=EYN2Md-DCR8

The modular pipeline



NER (Named Entity Recognition)



4 *Journal of the American Medical Informatics Association*, 2019, Vol. 00, No. 0

Table 1. Named entities and attributes recognized by Criteria2Query

	Category	Definition	Examples
Entity	Condition	Conditions are records of a Person suggesting the presence of a disease or medical condition stated as a diagnosis, a sign or a symptom.	*Type 2 diabetes mellitus, Alzheimer's disease.*
	Drug	Drugs are biochemical substances formulated in such ways that when administered to a person it will exert a certain physiological effect.	*Acetaminophen, Furosemide*
	Measurement	The standardized examination or testing of a person or person's sample.	*Serum creatinine, Serum bilirubin*
	Procedure	Procedures are activities or processes on the patient to have a diagnostic or therapeutic purpose.	*Chemotherapy, Radiotherapy*
	Observation	Observations are clinical facts about a person obtained in the context of examination, questioning or a procedure.	*Smoking, drug allergy*
Attribute	Value	Numeric attributes include but not limited to age range, lab test result, etc.	*30 to 75 years old*
	Temporal	Temporal constraints imposed on clinical diagnoses, drugs, etc.	*within 12 months*

Relations Extraction

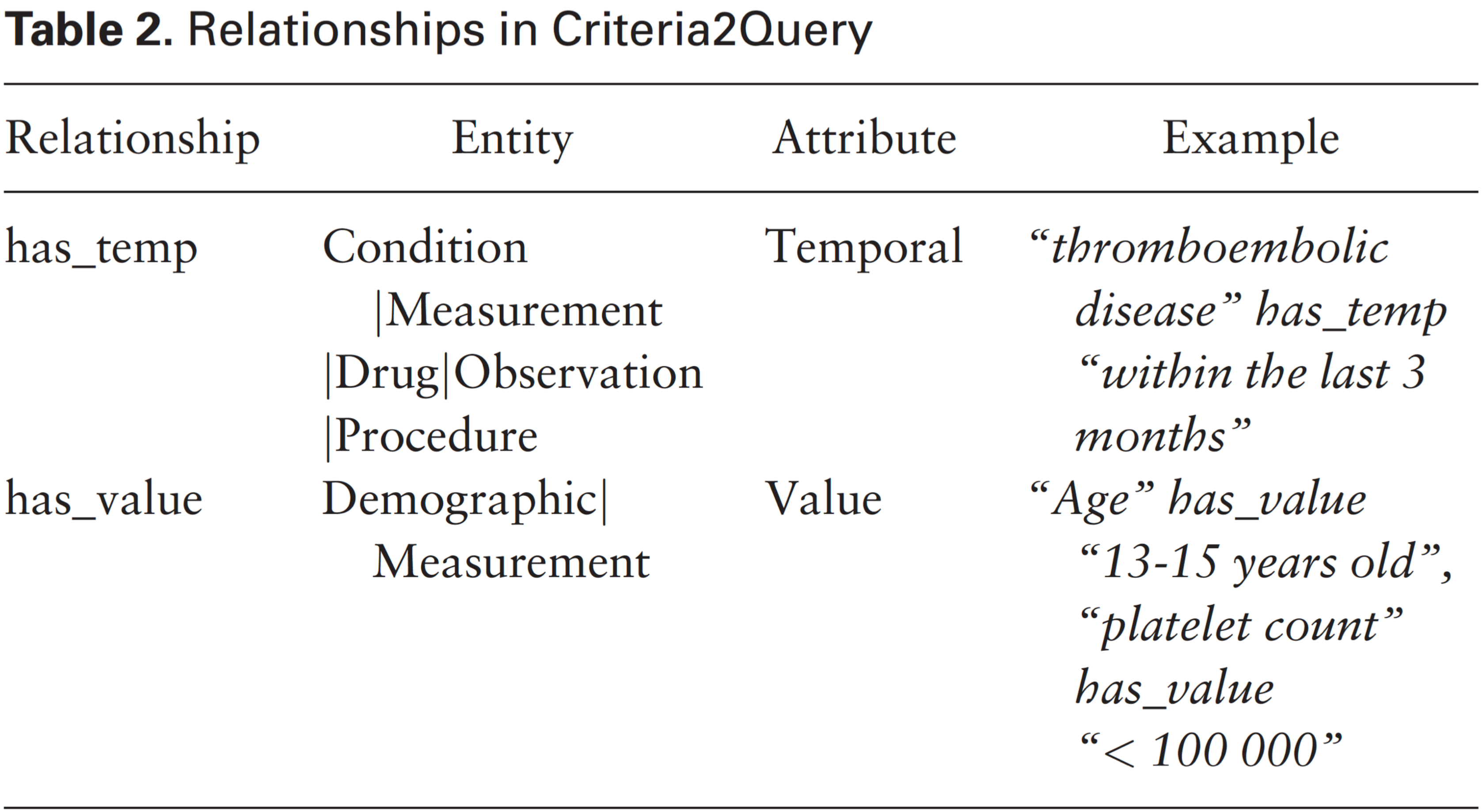
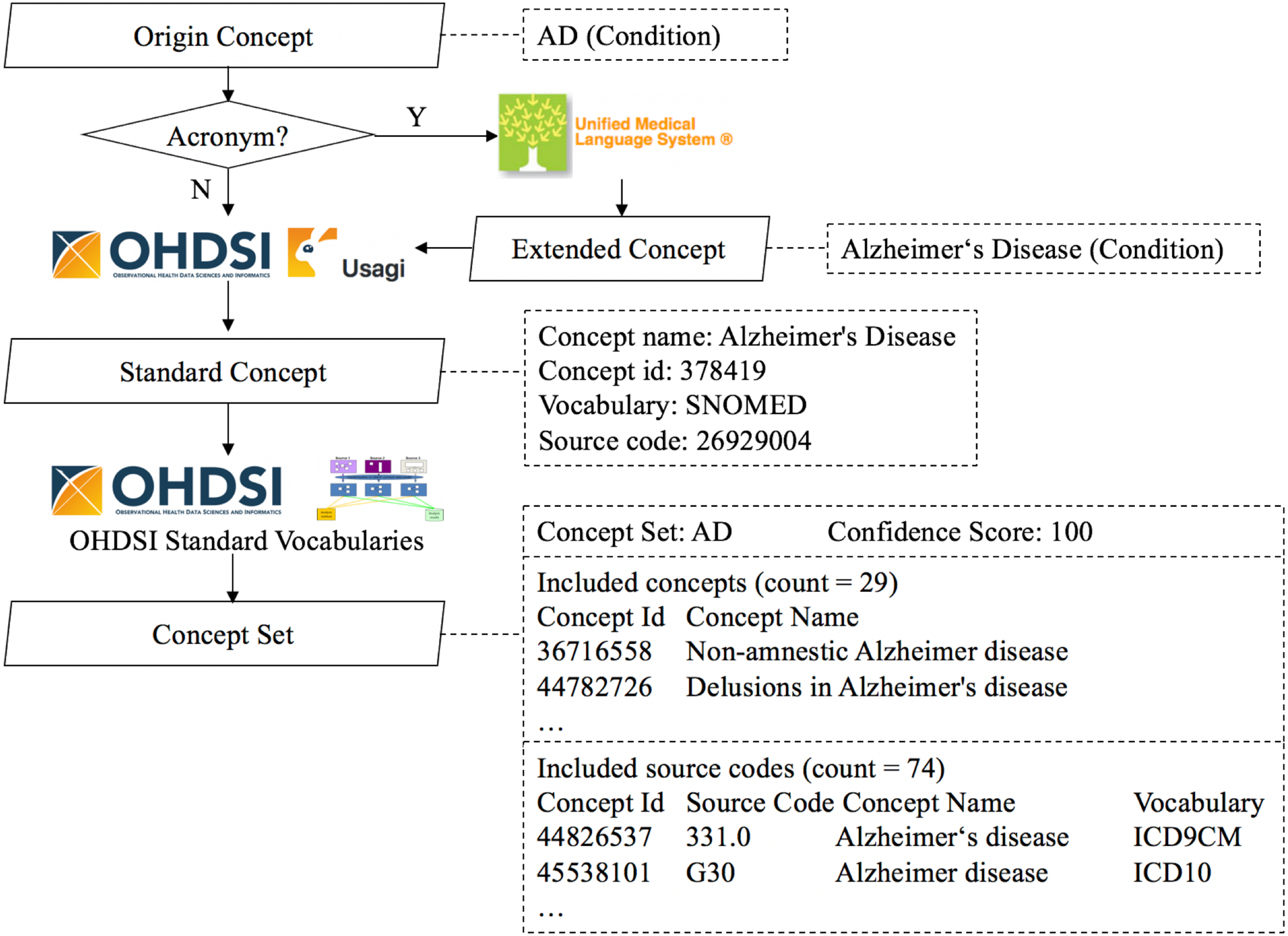


Table 2. Relationships in Criteria2Query

Relationship	Entity	Attribute	Example
has_temp	Condition |Measurement |Drug|Observation |Procedure	Temporal	*"thromboembolic disease" has_temp "within the last 3 months"*
has_value	Demographic| Measurement	Value	*"Age" has_value "13-15 years old", "platelet count" has_value "< 100 000"*

Entity normalization



Origin Concept
AD (Condition)
Acronym?
Y
Unified Medical Language System ®
N
OHDSI
Observational Health Data Sciences and Informatics
Usagi
Extended Concept
Alzheimer‘s Disease (Condition)
Standard Concept
OHDSI
Observational Health Data Sciences and Informatics
OHDSI Standard Vocabularies
Concept Set

Concept name: Alzheimer's Disease
Concept id: 378419
Vocabulary: SNOMED
Source code: 26929004

Concept Set: AD Confidence Score: 100

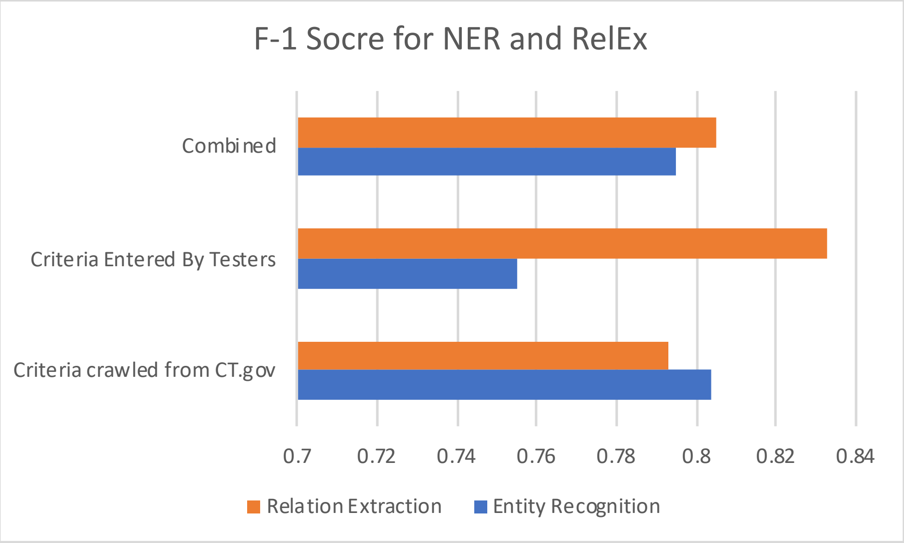
Included concepts (count = 29)

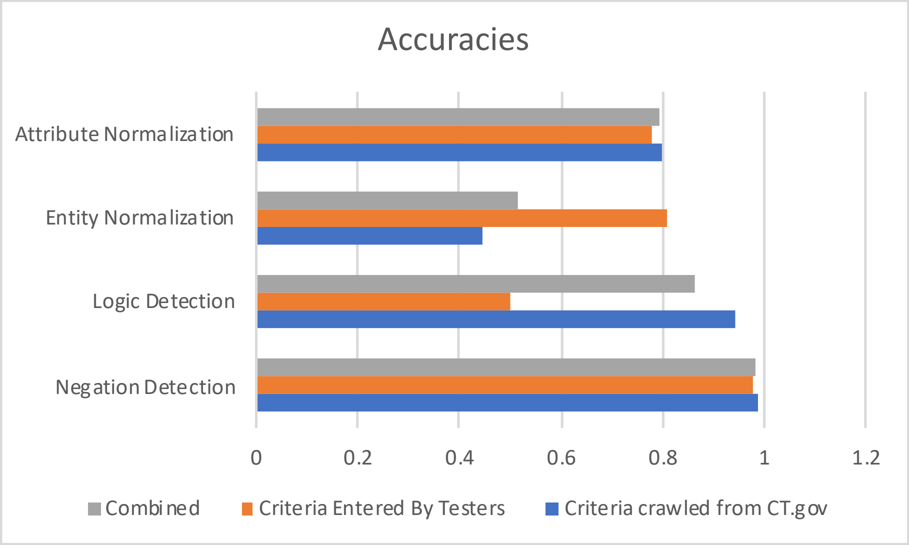
Concept Id	Concept Name
36716558	Non-amnestic Alzheimer disease
44782726	Delusions in Alzheimer's disease
…	

Included source codes (count = 74)

Concept Id	Source Code	Concept Name	Vocabulary
44826537	331.0	Alzheimer‘s disease	ICD9CM
45538101	G30	Alzheimer disease	ICD10
…			

Evaluation





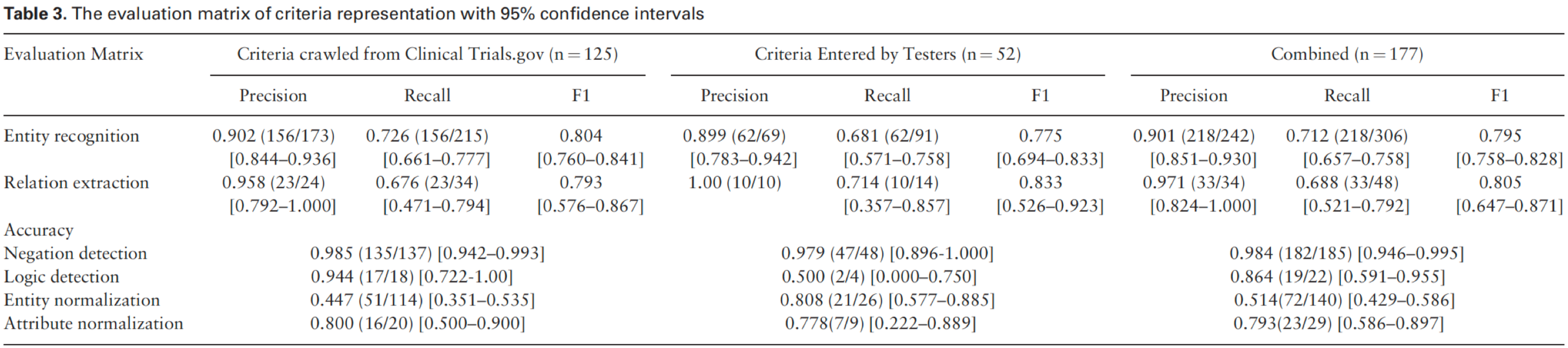


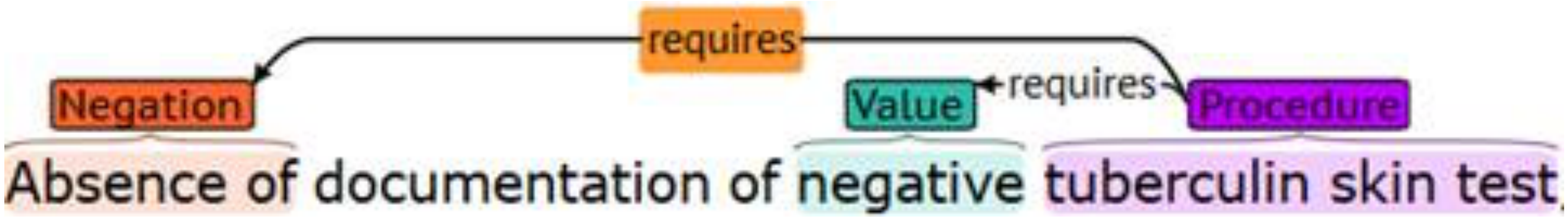
Table 3. The evaluation matrix of criteria representation with 95% confidence intervals

Evaluation Matrix	Criteria crawled from Clinical Trials.gov (n = 125)			Criteria Entered by Testers (n = 52)			Combined (n = 177)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Entity recognition	0.902 (156/173) [0.844–0.936]	0.726 (156/215) [0.661–0.777]	0.804 [0.760–0.841]	0.899 (62/69) [0.783–0.942]	0.681 (62/91) [0.571–0.758]	0.775 [0.694–0.833]	0.901 (218/242) [0.851–0.930]	0.712 (218/306) [0.657–0.758]	0.795 [0.758–0.828]
Relation extraction	0.958 (23/24) [0.792–1.000]	0.676 (23/34) [0.471–0.794]	0.793 [0.576–0.867]	1.00 (10/10)	0.714 (10/14) [0.357–0.857]	0.833 [0.526–0.923]	0.971 (33/34) [0.824–1.000]	0.688 (33/48) [0.521–0.792]	0.805 [0.647–0.871]
Accuracy									
Negation detection	0.985 (135/137) [0.942–0.993]			0.979 (47/48) [0.896-1.000]			0.984 (182/185) [0.946–0.995]		
Logic detection	0.944 (17/18) [0.722-1.00]			0.500 (2/4) [0.000–0.750]			0.864 (19/22) [0.591–0.955]		
Entity normalization	0.447 (51/114) [0.351–0.535]			0.808 (21/26) [0.577–0.885]			0.514(72/140) [0.429–0.586]		
Attribute normalization	0.800 (16/20) [0.500–0.900]			0.778(7/9) [0.222–0.889]			0.793(23/29) [0.586–0.897]		

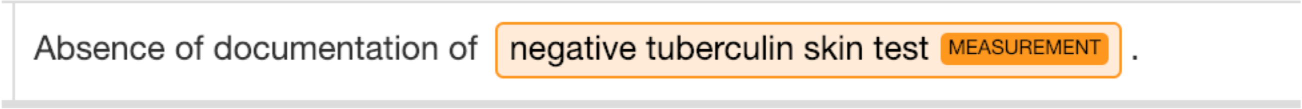
Error Analysis

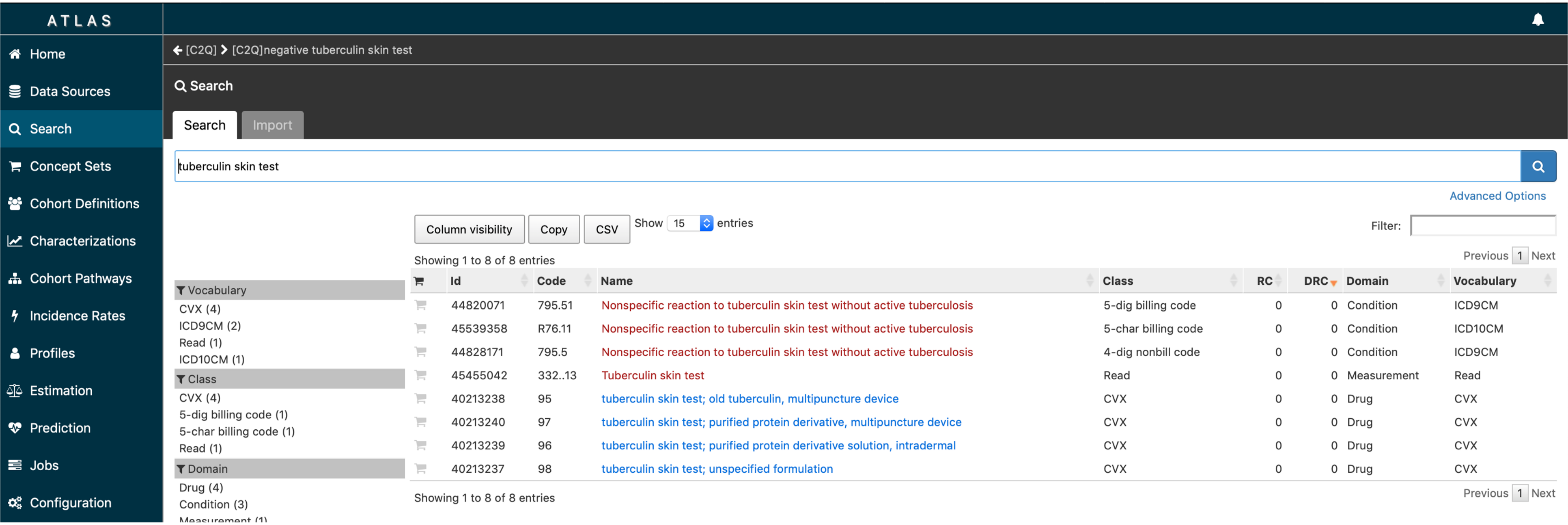
- Imperfect Information extraction results (NER, RelEx, Negation detection)
- Lack of medical knowledge, e.g., anti-inflammatory drugs, for concept normalization
- Incomplete concept coverage in OMOP CDM

Example Errors



Absence of documentation of negative tuberculin skin test


Absence of documentation of negative tuberculin skin test MEASUREMENT .



tuberculin skin test

Show 15 entries

Vocabulary
CVX (4)
ICD9CM (2)
Read (1)
ICD10CM (1)

Class
CVX (4)
5-dig billing code (1)
5-char billing code (1)
Read (1)

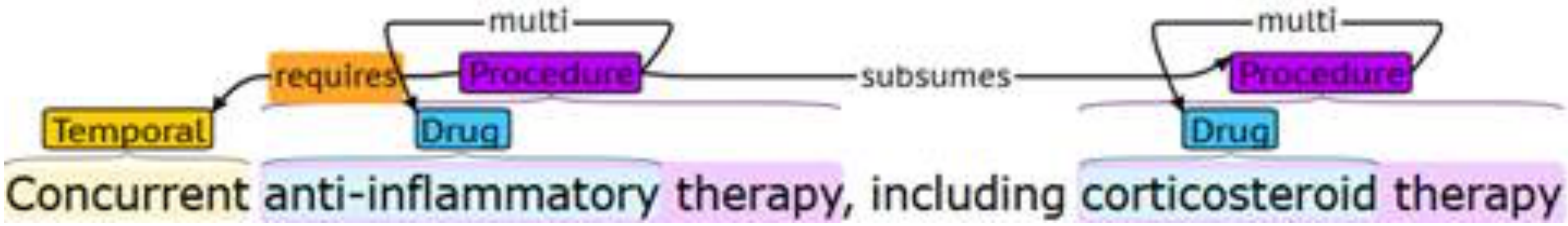
Domain
Drug (4)
Condition (3)
Measurement (1)

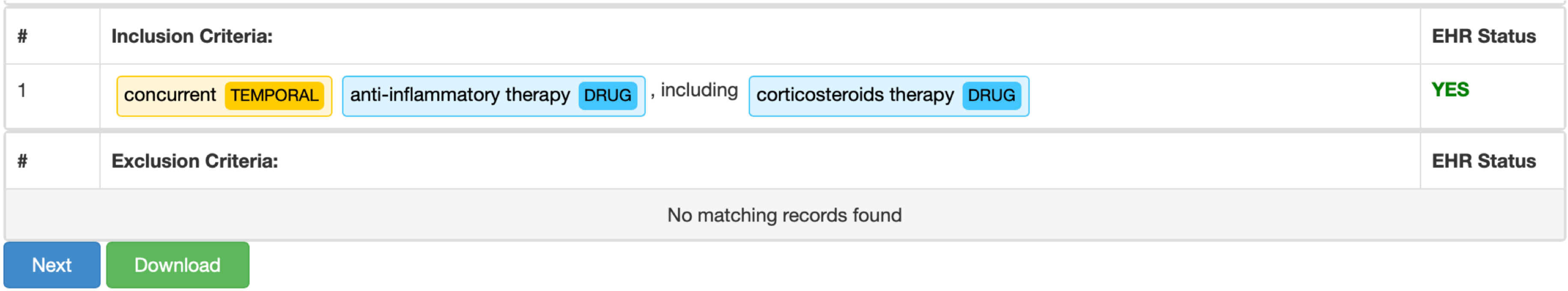
Showing 1 to 8 of 8 entries

Id	Code	Name	Class	RC	DRC	Domain	Vocabulary
44820071	795.51	Nonspecific reaction to tuberculin skin test without active tuberculosis	5-dig billing code	0	0	Condition	ICD9CM
45539358	R76.11	Nonspecific reaction to tuberculin skin test without active tuberculosis	5-char billing code	0	0	Condition	ICD10CM
44828171	795.5	Nonspecific reaction to tuberculin skin test without active tuberculosis	4-dig nonbill code	0	0	Condition	ICD9CM
45455042	332..13	Tuberculin skin test	Read	0	0	Measurement	Read
40213238	95	tuberculin skin test; old tuberculin, multipuncture device	CVX	0	0	Drug	CVX
40213240	97	tuberculin skin test; purified protein derivative, multipuncture device	CVX	0	0	Drug	CVX
40213239	96	tuberculin skin test; purified protein derivative solution, intradermal	CVX	0	0	Drug	CVX
40213237	98	tuberculin skin test; unspecified formulation	CVX	0	0	Drug	CVX

Showing 1 to 8 of 8 entries

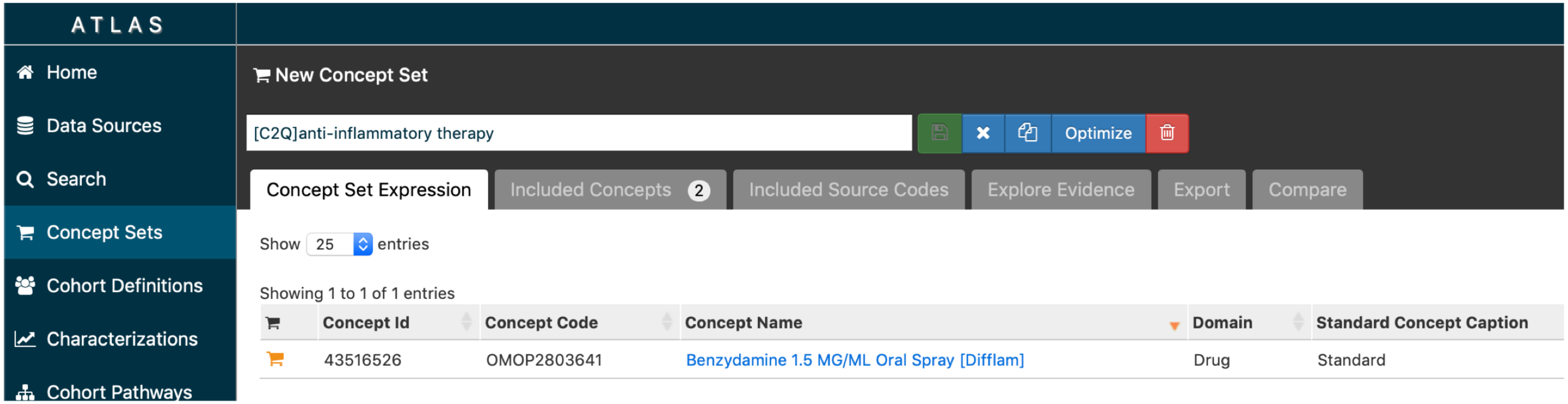
Example Errors


multi
requires
Procedure
subsumes
multi
Procedure
Temporal
Drug
Drug
Concurrent anti-inflammatory therapy, including corticosteroid therapy



#	Inclusion Criteria:	EHR Status
1	concurrent TEMPORAL anti-inflammatory therapy DRUG , including corticosteroids therapy DRUG	YES

#	Exclusion Criteria:	EHR Status
	No matching records found	



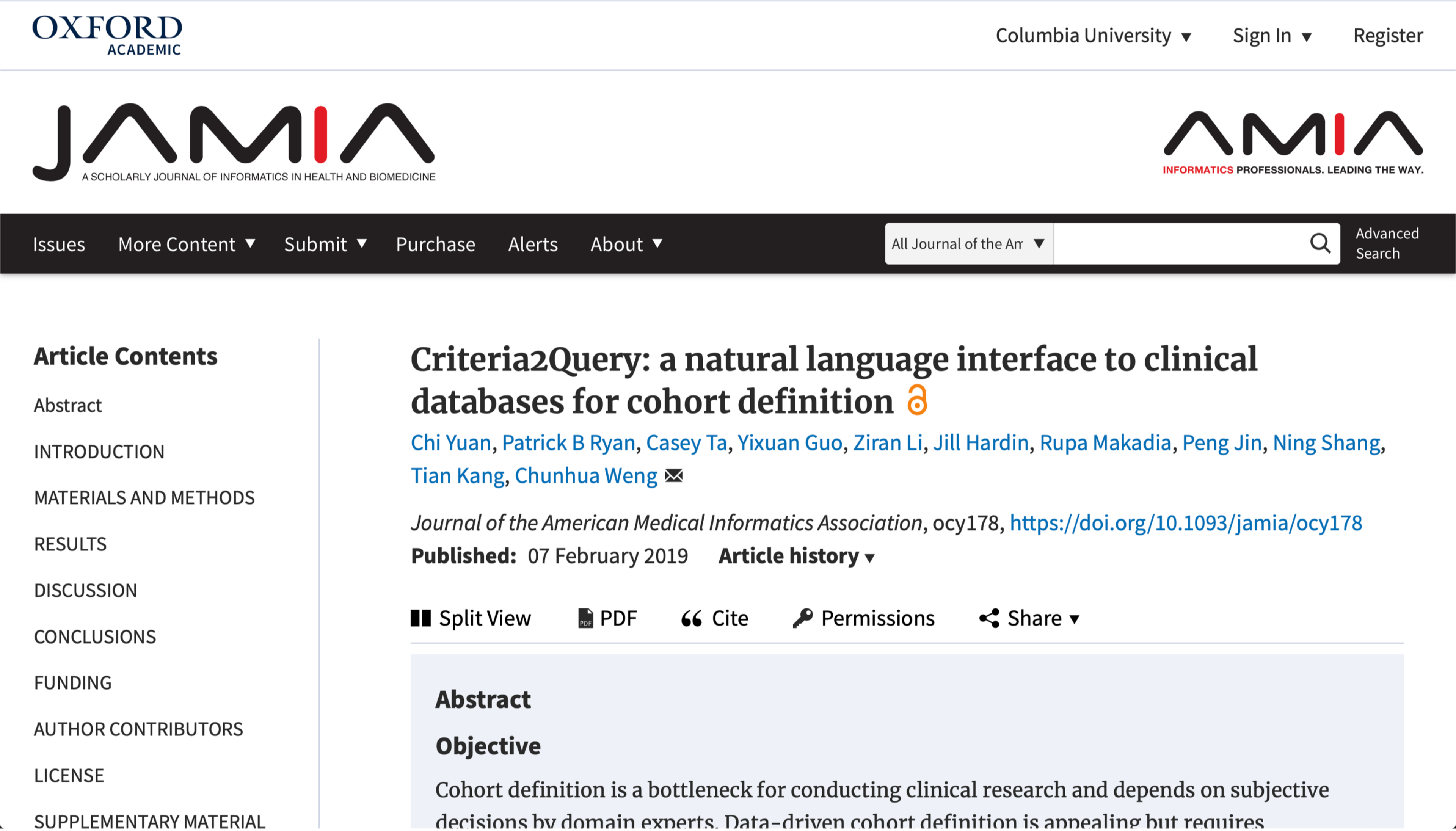
[C2Q]anti-inflammatory therapy

Show 25 entries

Showing 1 to 1 of 1 entries

	Concept Id	Concept Code	Concept Name	Domain	Standard Concept Caption
	43516526	OMOP2803641	Benzydamine 1.5 MG/ML Oral Spray [Difflam]	Drug	Standard

Preliminary Progress



Criteria2Query: a natural language interface to clinical databases for cohort definition

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, Chunhua Weng ✉

Journal of the American Medical Informatics Association, ocy178, https://doi.org/10.1093/jamia/ocy178

Published: 07 February 2019 **Article history** ▾

Abstract

Objective

Cohort definition is a bottleneck for conducting clinical research and depends on subjective decisions by domain experts. Data-driven cohort definition is appealing but requires

https://doi.org/10.1093/jamia/ocy178

Open source resources

- Introduction https://www.youtube.com/watch?v=EYN2Md-DCR8
- Open source: https://github.com/OHDSI/Criteria2Query
- Online system: http://www.ohdsi.org/web/criteria2query/
- Feedback or inquiries: https://gitter.im/Criteria2query/Lobby#

Contributions

- An early natural language interface to clinical database
- An open-source pipeline with modular architecture

Ongoing work for collaboration

- Richer annotated corpus of criterion text
- State of the art NLP methods application
- More intelligent concept set recommendation
- More user-friendly interactive design

Ongoing projects

- Mapping of Note Types to LOINC/standard vocabulary – Karthik Natarajan, Ruth Reeves, and Jon Duke
- Landscape Analysis of section identifier systems and proposal of a standard terminology for use – Hua Xu and Karthik Natarajan
- Mapping of CUIs to standard terminology – Juan Banda
- Standardization of term_modifiers and values – Hua Xu
- Evaluate and revise textual CDM tables by sharing practical issues and lessons learnt during ETL for processing textual data into CDM – Ruth Reeves, others?
- Develop tools (within Atlas) to facilitate uses of NLP data for cohort building/phenotyping : Collaborate with eMERGE consortium:
- Conduct cross-site studies that use textual data
- Continue developing other NLP resources

Other issues

- Meeting formats : Presentation followed by updates on ongoing projects
- Presentation scheduling
  - March 13th – Yuan Luo – eMERGE collaboration
  - April 10th – Jon Duke – ClarityNLP
  - May 8th – Juan Banda - CUI mapping, ongoing work – Juan, Stephan Meyestre – tool to evaluate NLP systems
  - June 12th
  - July 10th
- Please let us know if you can present your related work at any of the above meetings.