Feature Engineering to Power Machine Learning Phenotype Development

PRESENTER:
Xinzhuo Jiang, Krishna Kalluri, Chao Pang

INTRO:

In order to quickly discover new phenotypes, we leveraged state-of-the-art phenotype definitions as the gold standard to train ML models for predicting pairs of concepts that could potentially belong together to the same phenotype.

METHODS:

We created a training set (see **Table.1**) with positive and negative pairs of concepts extracted from the gold standard, then applied 3 distinct techniques to generate 21 features for a machine learning (ML) group to use for training. Our extensible feature engineering pipeline was designed to run on various data sources.

1. Lexical features

*Measure the degree to which the two concepts are lexically similar, see **Fig.1***

2. Data-driven features

- co-occurrence matrix

*Compute the relative frequency of two events co-occurring within the specific time window, see **Fig.2***

3. Knowledge-based features

- semantic similarity

*Compute the likeness of their meaning or semantic content, see **Fig.3***

RESULTS:

- The high-throughput feature engineering approach provides ready-to-use features for similar ML problems. In addition, the feature pipeline can be run on OMOP directly or other data sources with minor tweaks.
- The feature importance scores show that knowledge representation is more significant than data driven and lexical features in terms of the prediction power for this specific ML problem.

A **high-throughput** feature engineering approach for phenotyping in OMOP

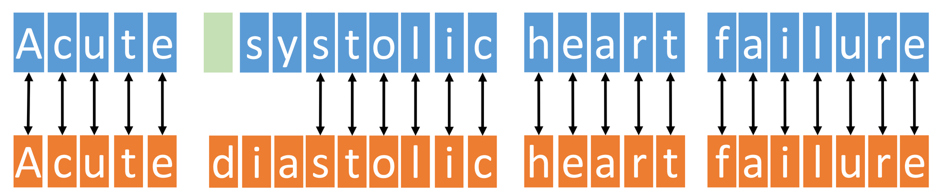
Take a picture to download the full paper

Table.1 Data source

ground_truth	concept_id_1	concept_id_2	concept_name_1	concept_name_2	same_domain	is_ancestor	min_distance
0	75576	435216	Irritable bowel syndrome	Disorder due to type 1 diabetes mellitus	1	0	9
0	80809	378726	Rheumatoid arthritis	Dementia associated with alcoholism	1	0	6
1	81064	439770	Pseudopolyposis of colon	Ketoacidosis in type 1 diabetes mellitus	1	0	9
0	81097	26942	Felty's syndrome	Hemoglobin SS disease with crisis	1	0	6
1	81097	1119155	Felty's syndrome	0.4 ML adalimumab 50 MG/ML Prefilled Syringe [Humira]	0	0	

Example
Concept_1: **Acute systolic heart failure**
Concept_2: **Acute diastolic heart failure**

Fig.1 Levenshtein_ratio



$$Ratio = \frac{levenshtein\_distance(concept\_1, concept\_2)}{max(len(concept\_1), len(concept\_2))}$$

Fig.2 Co-occurrence matrix in 180 days

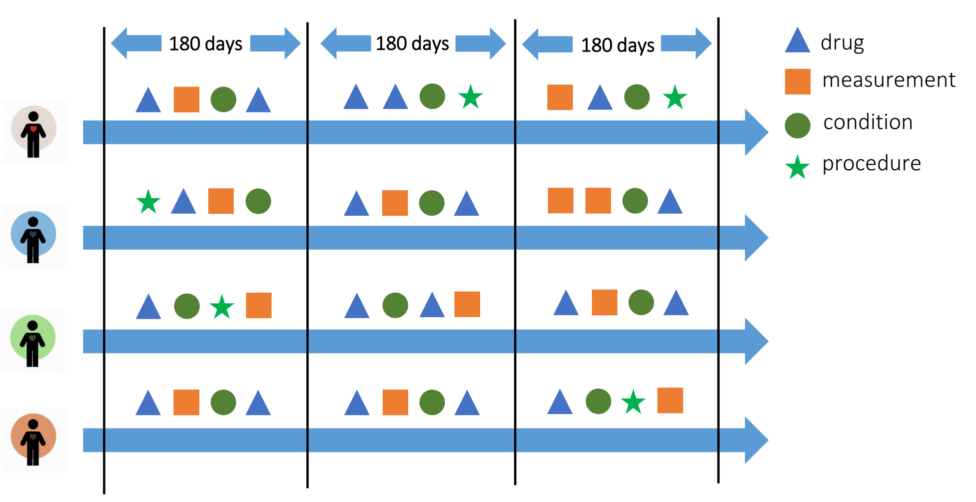
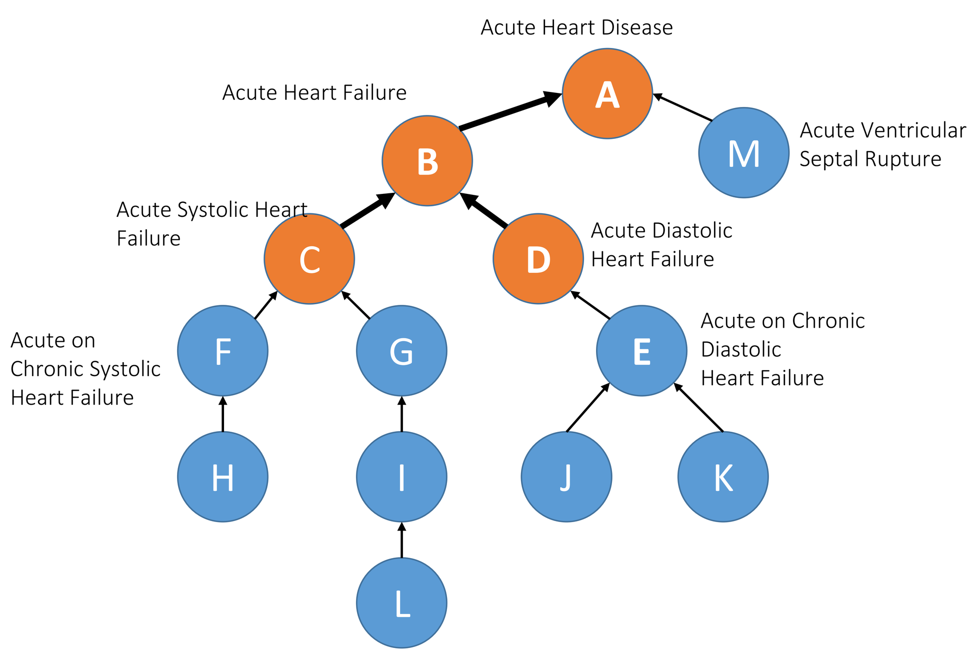


Fig.3 Semantic similarity



$$Similarity\ (C, D) = \frac{2 * len(A,B)}{(len(A,C) + len(A,D))}$$

Table.2 Available features

Lexical Features	Data-driven Features	Knowledge-based Features
levenshtein_distance	cooccurrence_visit	distance_indicator
levenshtein_ratio	cooccurrence_60_days	information_content
jaro	cooccurrence_180_days	semantic_similarity
jaro_winkler	cooccurrence_360_days	lin_measure
fuzz_partial_ratio	cooccurrence_lifetime	jiang_measure
	lifetime_cooccur_embedding_cosine	relevance_measure
	5_year_cooccur_embedding_cosine	information_coefficient
	visit_cooccur_embedding_cosine	graph_ic_measure

Xinzhuo Jiang, Krishna Kalluri, Chao Pang, Kai Chen, Junghwan Lee , Cong Liu, Ruijun Chen, Patrick Ryan, Karthik Natarajan

Columbia University Medical Center
OHDSI