A smart tool for query and risk calculation

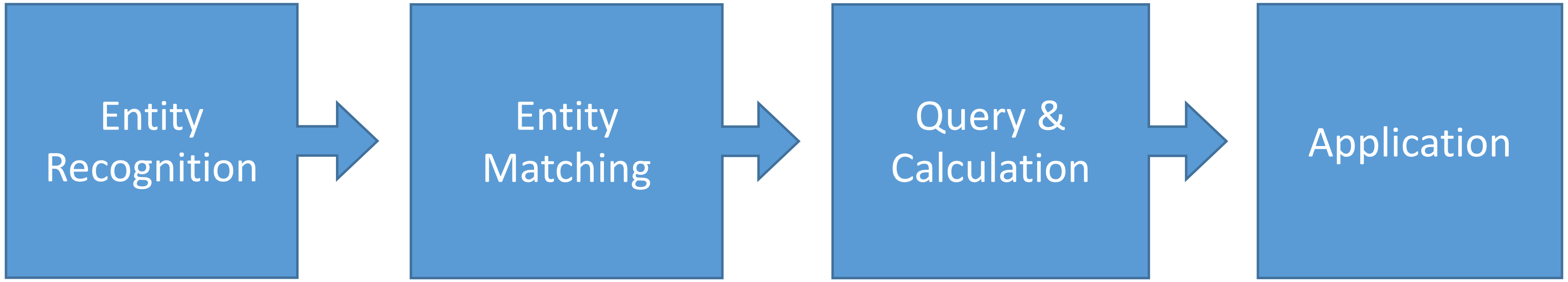
Feiyi Wang

Supervisor: Andrea Ganna

Introduction

- A tool for users
    - To query the risk of a disease given another disease
    - To understand how to query the original Risteys R6 data given a specific question
    - To understand how to culculate the risk using the data
- An application which can also understand users' questions well

Workflow



Entity Recognition

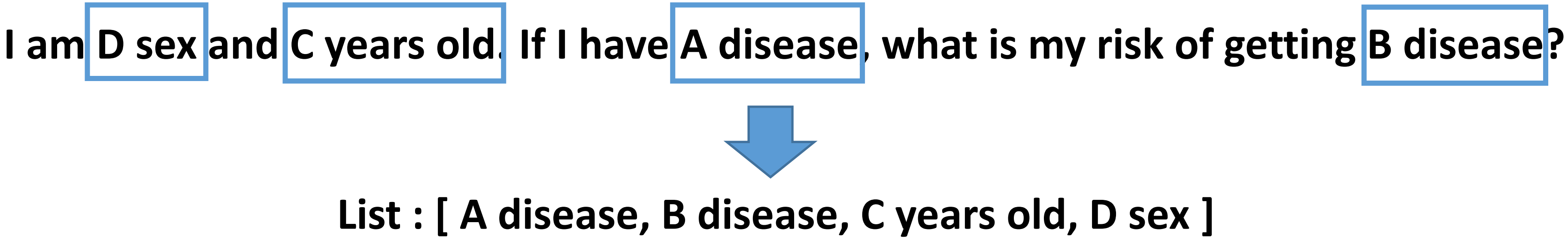
Entity Matching

Query & Calculation

Application

Entity Recognition - Extract keywords from a question

Unstructured text :



I am D sex and C years old. If I have A disease, what is my risk of getting B disease?

List : [A disease, B disease, C years old, D sex]

Questions	Answers
What is my risk of angina if I am a female with a history of heart attack?	['heart attack', 'angina', 'na', 'female]
What's the risk of having cancer if I have heart attavk?	['heart attack', 'cancer', 'na', 'na']
If I am a girl at 24 with cancer, what's my risk of having diabeties?	['cancer', 'diabetes', '24', 'female']
I am a 65-year-old male. I had strke. What is my risk of epilepsy?	['stroke', 'epilepsy', '65', 'male']
I am 20 and I have cancer	['cancer', 'na', '20', 'na']

GPT-3

- Generative Pre-trained Transformer 3 (GPT-3) is an autoregressive language model that uses deep learning to produce human-like text [1]

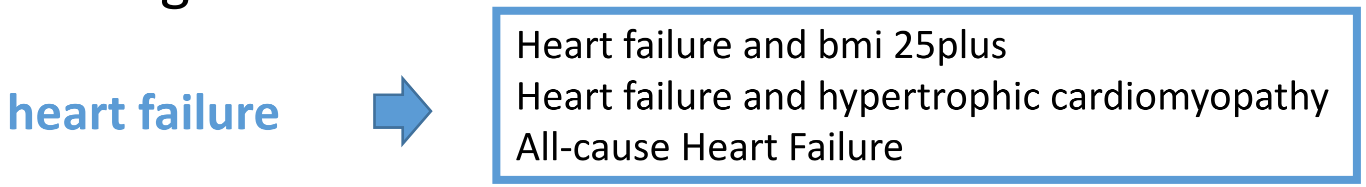
Questions	Answers	davinci	curie	babbage	ada
What is my risk of angina if I am a female with a history of heart attack?	['heart attack', 'angina', 'na', 'female]	['heart attack', 'angina', 'na', 'female]	['heart attack', 'angina', 'na', 'female]	['heart attack', 'angina', 'female', 'na']	['heart attack', 'angina', 'na', 'female']
What's the risk of having cancer if I have heart attavk?	['heart attack', 'cancer', 'na', 'na']	['cancer', 'heart attack', 'na', 'na']	['heart attack', 'cancer', 'na', 'na']	['cancer', 'heart attavk', 'na', 'na']	['heart attavk', 'cancer', 'na', 'na']
If I am a girl at 24 with cancer, what's my risk of having diabeties?	['cancer', 'diabetes', '24', 'female']	['cancer', 'diabetes', '24', 'female']	['cancer', 'diabeties', '24', 'female']	['cancer', 'diabetes', '24', 'female']	['cancer', 'diabeties', 'na', 'na']
I am a 65-year-old male. I had strke. What is my risk of epilepsy?	['stroke', 'epilepsy', '65', 'male']	['stroke', 'epilepsy', '65', 'male']	['stroke', 'epilepsy', '65', 'male']	['epilepsy', 'stroke', '65', 'male']	['epilepsy', 'stroke', 'na', 'na']
I am 20 and I have cancer	['cancer', 'na', '20', 'na']	['cancer', '20', 'na']	['cancer', '20', 'na']	['cancer', 'na', '20', 'female']	['cancer', '20', 'male']

- Engine selection:
  - if all elements can be found in the sentence
  - if the order of disease names can be recognized
  - if misspelling can be detected and fixed

[1] https://en.wikipedia.org/wiki/GPT-3

Entity Matching - Find pairwise endpoints

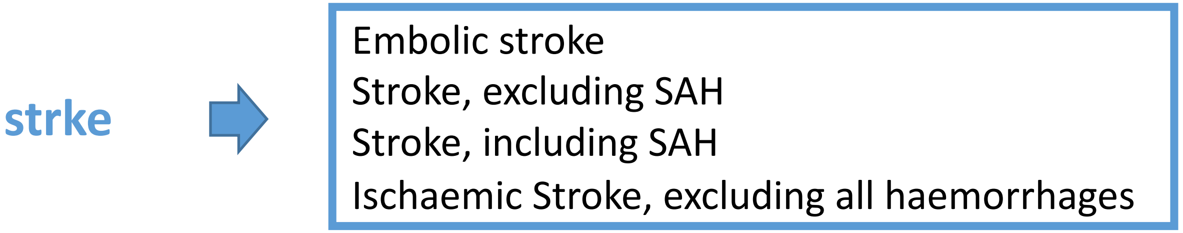
- Keyword exact matching



heart failure

Heart failure and bmi 25plus
Heart failure and hypertrophic cardiomyopathy
All-cause Heart Failure

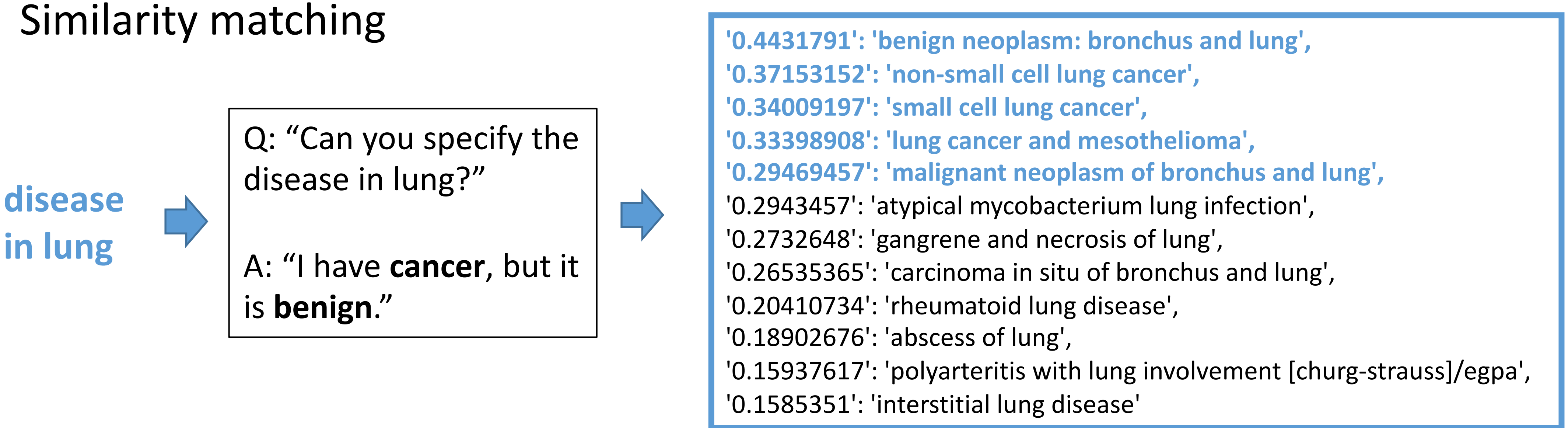
- Fuzzy matching



strke

Embolic stroke
Stroke, excluding SAH
Stroke, including SAH
Ischaemic Stroke, excluding all haemorrhages

- Similarity matching



disease in lung

Q: “Can you specify the disease in lung?”

A: “I have **cancer**, but it is **benign**.”

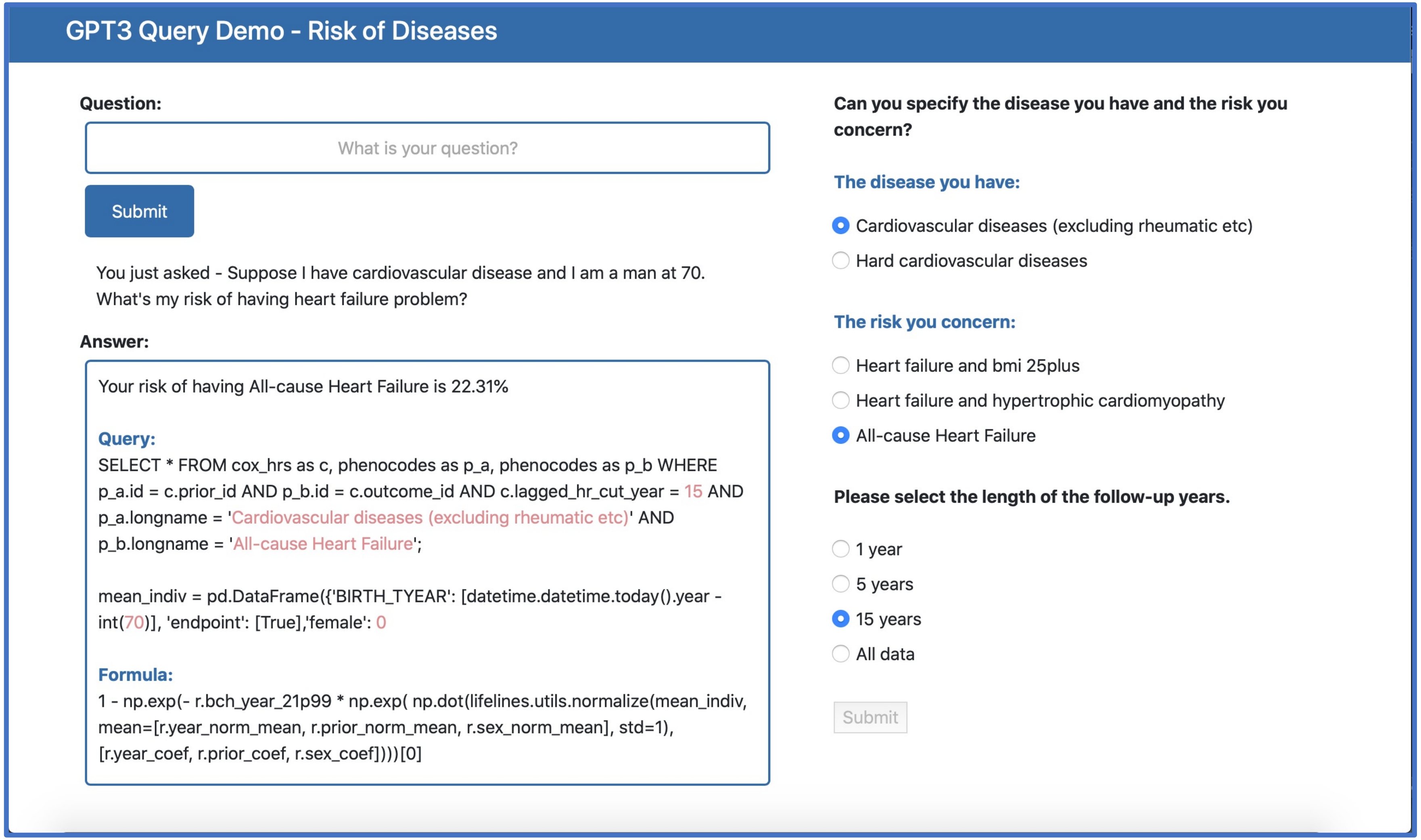
'0.4431791': 'benign neoplasm: bronchus and lung',
'0.37153152': 'non-small cell lung cancer',
'0.34009197': 'small cell lung cancer',
'0.33398908': 'lung cancer and mesothelioma',
'0.29469457': 'malignant neoplasm of bronchus and lung',
'0.2943457': 'atypical mycobacterium lung infection',
'0.2732648': 'gangrene and necrosis of lung',
'0.26535365': 'carcinoma in situ of bronchus and lung',
'0.20410734': 'rheumatoid lung disease',
'0.18902676': 'abscess of lung',
'0.15937617': 'polyarteritis with lung involvement [churg-strauss]/egpa',
'0.1585351': 'interstitial lung disease'

Query & Formula - Risk calculation

- Collect the birth year and sex from user
- Query
  - Mean of birth year
  - Mean of sex: 1 if female; 0 otherwise
  - Incidence of the prior disease
- Normalize the collected data
- Formula

$$1 - e^{BaselineCumulativeHazard \times e^{NormalizedData \cdot Coef}}$$

Application- Demo



GPT3 Query Demo - Risk of Diseases

Question:

What is your question?

Submit

You just asked - Suppose I have cardiovascular disease and I am a man at 70. What's my risk of having heart failure problem?

Answer:

Your risk of having All-cause Heart Failure is 22.31%

Query:

```
SELECT * FROM cox_hrs as c, phenocodes as p_a, phenocodes as p_b WHERE
p_a.id = c.prior_id AND p_b.id = c.outcome_id AND c.lagged_hr_cut_year = 15 AND
p_a.longname = 'Cardiovascular diseases (excluding rheumatic etc)' AND
p_b.longname = 'All-cause Heart Failure';

mean_indiv = pd.DataFrame({'BIRTH_TYEAR': [datetime.datetime.today().year -
int(70)], 'endpoint': [True],'female': 0
```

Formula:

```
1 - np.exp(- r.bch_year_21p99 * np.exp( np.dot(lifelines.utils.normalize(mean_indiv,
mean=[r.year_norm_mean, r.prior_norm_mean, r.sex_norm_mean], std=1),
[r.year_coef, r.prior_coef, r.sex_coef])))[0]
```

Can you specify the disease you have and the risk you concern?

The disease you have:

● Cardiovascular diseases (excluding rheumatic etc)

○ Hard cardiovascular diseases

The risk you concern:

○ Heart failure and bmi 25plus

○ Heart failure and hypertrophic cardiomyopathy

● All-cause Heart Failure

Please select the length of the follow-up years.

○ 1 year

○ 5 years

● 15 years

○ All data

Submit

Next Steps

- More queries and formulas
  - e.g. How many **women** below **65** have been diagnosed as **coronary artery disease** between **2015** and **2018**?
- Accuracy of the entity recognition/matching algorithm
  - e.g. More accurately capture the order of disease and risk
- Better exception handling
- Better user experience

Acknowledgements

- Thank you to Andrea Ganna & Vincent Llorens