

THE DATA INCUBATOR – PROJECT PROPOSAL

Data Mining from Clinical Research Records for Healthcare

Jie Yang

Ph.D. Student at Department of Biomedical Engineering, Columbia University

Research Area: Machine Learning for Medical Image Classification / Segmentation for Healthcare

Introduction

<https://clinicaltrials.gov/>



ClinicalTrials.gov

Find Studies ▾ About Studies ▾ Submit Studies ▾ Resources ▾ About Site ▾

ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world.

5GB of text records

Explore 266,538 research studies in all 50 states and in 203 countries.

ClinicalTrials.gov is a resource provided by the U.S. National Library of Medicine.

IMPORTANT: Listing a study does not mean it has been evaluated by the U.S. Federal Government. Read our [disclaimer](#) for details.

Before participating in a study, talk to your health care provider and learn about the [risks and potential benefits](#).

Find a study (all fields optional)

Recruitment status i

- Recruiting and not yet recruiting studies
 All studies

Condition or disease i (For example: breast cancer)

Other terms i (For example: NCT number, drug name, investigator name)

Country i

Search

[Advanced Search](#)

Query for the disease

Query for location of study

Introduction

<https://clinicaltrials.gov/>

Rich information:

- Disease type;
- Study location;
- Proposed treatment;
- Detailed description;
- Starting date;
- Ongoing / final results; ...

For patient use

For research use

The screenshot shows the ClinicalTrials.gov website. At the top, it features the NIH logo and the text "U.S. National Library of Medicine". Below this, the "ClinicalTrials.gov" logo is displayed. A navigation bar with links to "Find Studies", "About Studies", "Submit Studies", "Resources", and "About Site" is visible. The main content area has a dark blue background with white text. It states: "ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world." Below this, there is a section titled "Explore 266,538 research studies in all 50 states and in 203 countries." It also includes a note about the resource being provided by the U.S. National Library of Medicine and a disclaimer regarding the evaluation of studies by the U.S. Federal Government. On the right side, there is a search interface titled "Find a study (all fields optional)". It includes fields for "Recruitment status" (with radio buttons for "Recruiting and not yet recruiting studies" and "All studies", where "All studies" is selected), "Condition or disease" (with a text input field and an "X" button to clear it), "Other terms" (with a text input field and an "X" button to clear it), and "Country" (with a dropdown menu and an "X" button to clear it). At the bottom of the search interface are two buttons: "Search" and "Advanced Search".

Proposal

Develop data science interface for mining this dataset:



Time consuming to query for the information.



Better better summarization and visualization of the data:

- E.g. to inspect ***past*** treatments, ***trending*** treatments, and predict ***future*** plans.

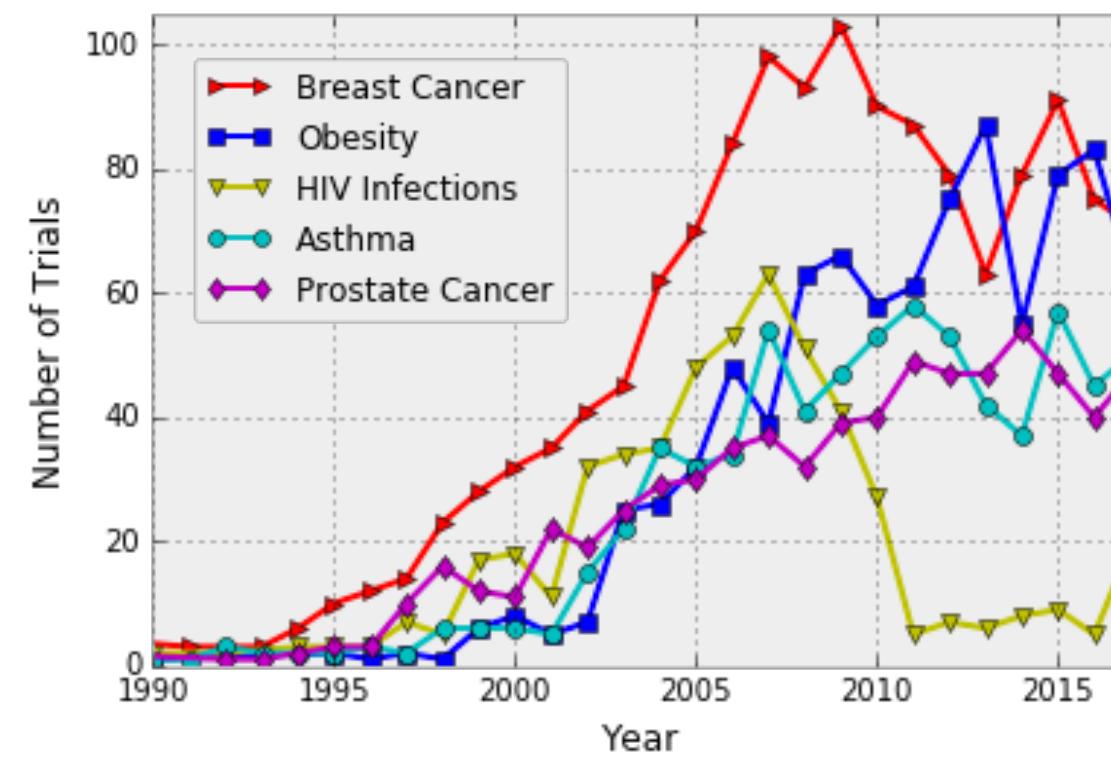
for
researcher

for
patient

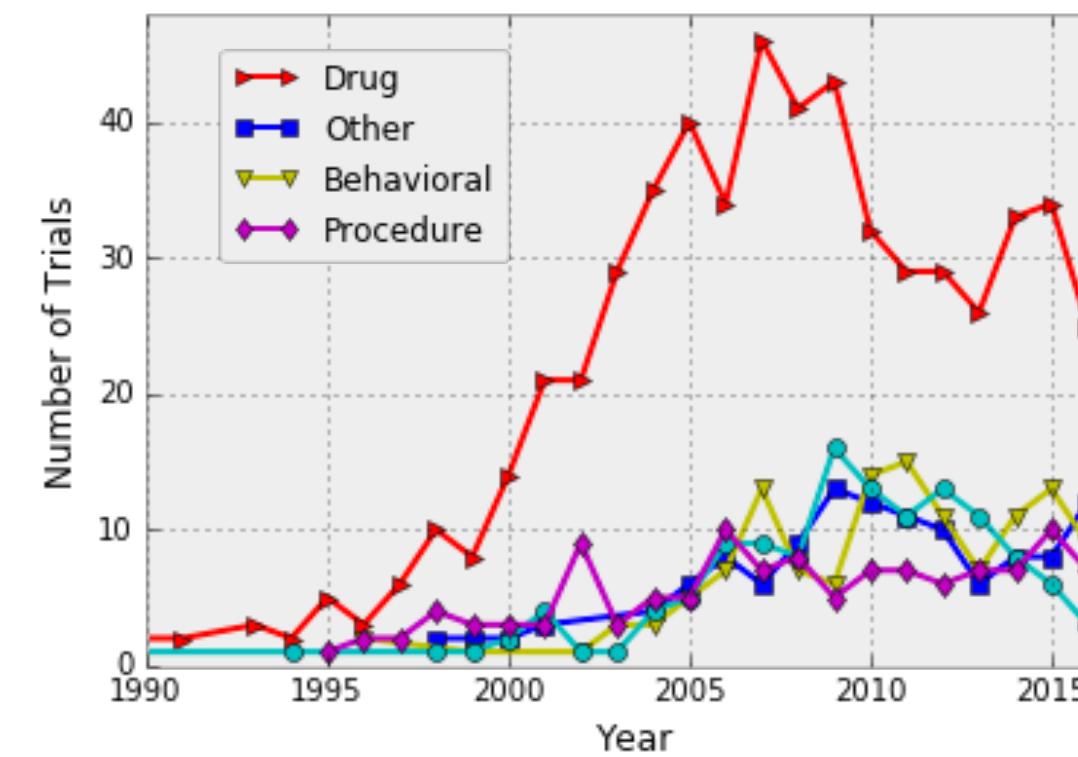
for
company

Proposal

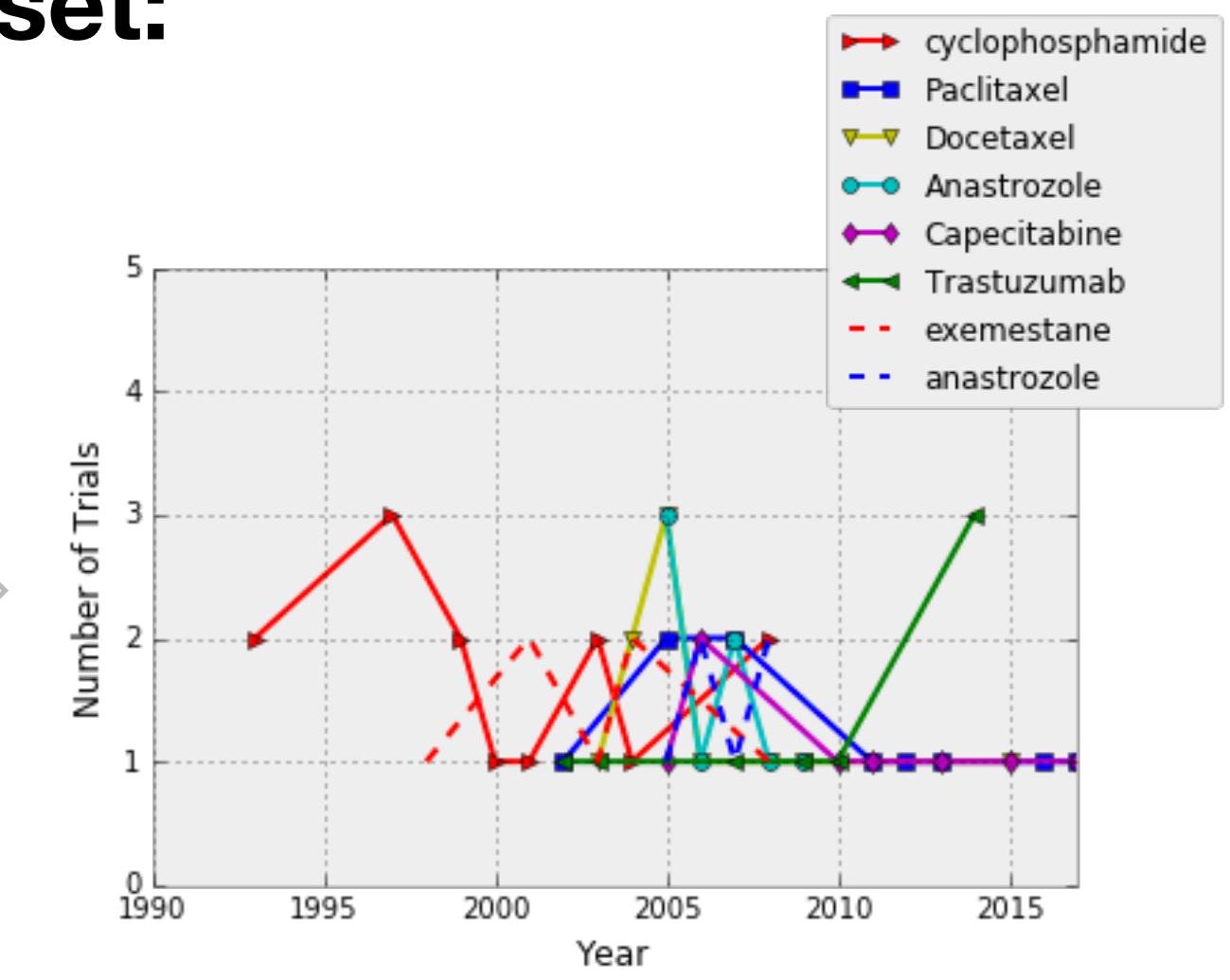
Develop data science interface for mining this dataset:



Popular Disease



Popular Treatment Type



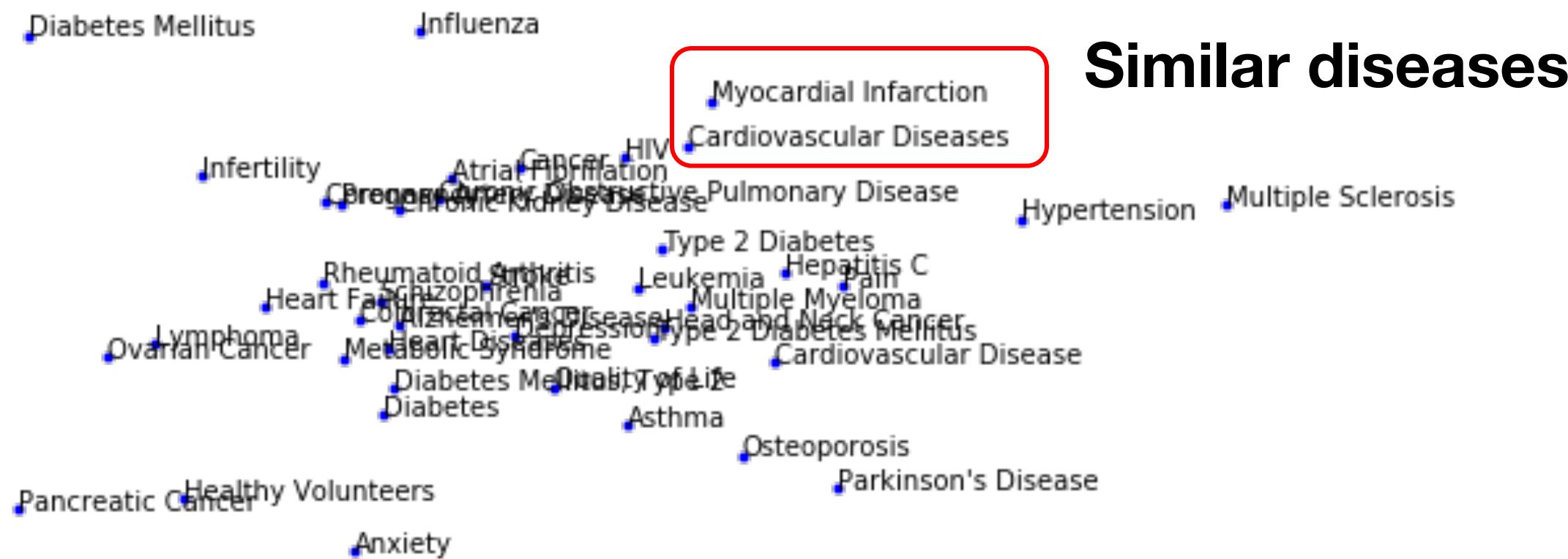
Popular Treatment Name

https://github.com/wanwanbeen/text_mining_ClinicalTrials/blob/master/demo_ClinicalTrials_explore.ipynb

Proposal

Develop data science interface for mining this dataset:

- ! Different study: Same / similar / correlated disease but registered with different names.
- 💡 Better better query with word embedding:



https://github.com/wanwanbeen/text_mining_ClinicalTrials/blob/master/demo_ClinicalTrials_embedding.ipynb

Proposal

Future plan:

- Learn better embedding from more detailed study description
- Categorize study results for treatment query
- Additional functions: e.g. study location map, ...