Improving Case Based Search Of Trials

Mentor : Soumyadeep Roy

Submitted By:

Mayank Jain

Pranjal Doshi

Udit Agarwal

Recap

NIH U.S. National Library of Medicine

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.

Explore 284,214 research studies in all 50 states and in 204 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)

Status

- ○ Recruiting and not yet recruiting studies
- ○ All studies

Condition or disease (For example: breast cancer)

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

Country

Search | Advanced Search

Help | Studies by Topic | Studies on Map | Glossary

Eligibility Criteria

Go to

Information from the National Library of Medicine NIH NLM

Choosing to participate in a study is an important personal decision. Talk with your doctor and family members or friends about deciding to join a study. To learn more about this study, you or your doctor may contact the study research staff using the contacts provided below. For general information, Learn About Clinical Studies.

Ages Eligible for Study:	20 Years and older (Adult, Older Adult)
Sexes Eligible for Study:	All
Sampling Method:	Non-Probability Sample

Study Population

outpatient, ward

Criteria

Inclusion Criteria:

- Patients older than 20 years diagnosed with interstitial lung disease
- Diagnostic criteria for interstitial lung disease. If one of the following is met:
    1. clinical suspicion of idiopathic pulmonary fibrosis (IPF); Characteristic chest CT findings with honeycomb cysts and fibrosis and reasonable clinical signs
    2. suspected interstitial pneumonia, or confirmed by biopsy with no evidence of infection : IPF, Non-specific interstitial pneumonia(NSIP), Cryptogenic organizing pneumonia(COP), unclassified fibrosis
    3. interstitial lung disease suspects with underlying rheumatic disease

Exclusion Criteria:

- No specific criteria

Contacts and Locations

Go to

Information from the National Library of Medicine NIH NLM

Study Description

Go to

Brief Summary:

In patients with interstitial lung disease (ILD) with inconsistent clinical and radiological features, establishing a reliable diagnosis of ILD requires a surgical lung biopsy

Transbronchial cryobiopsy is a minimally invasive, rapid, safe technique, and with histologic diagnostic yields, for ILD, typically exceeding 70 -80% .

The aim of this study is to compare and analyze the diagnostic yield, for ILD, and complications following SLB and TC

Methods. The investigators designed a descriptive, comparative and cross-sectional study in patients with ILD, in which SLB and CT will be performed in the same surgical stage, as diagnostic tests.

This study will be conducted from January 2018 to January 2019. Surgical lung biopsy and TC will be performed in the same surgical stage in all patients, under general anesthesia and mechanical ventilation. First TC will be performed by a pulmonologist, sequentially a thoracic surgeon will carry out a SLB.

The samples obtained will be analyzed by different pathologist to compare both techniques in terms of histologic features.

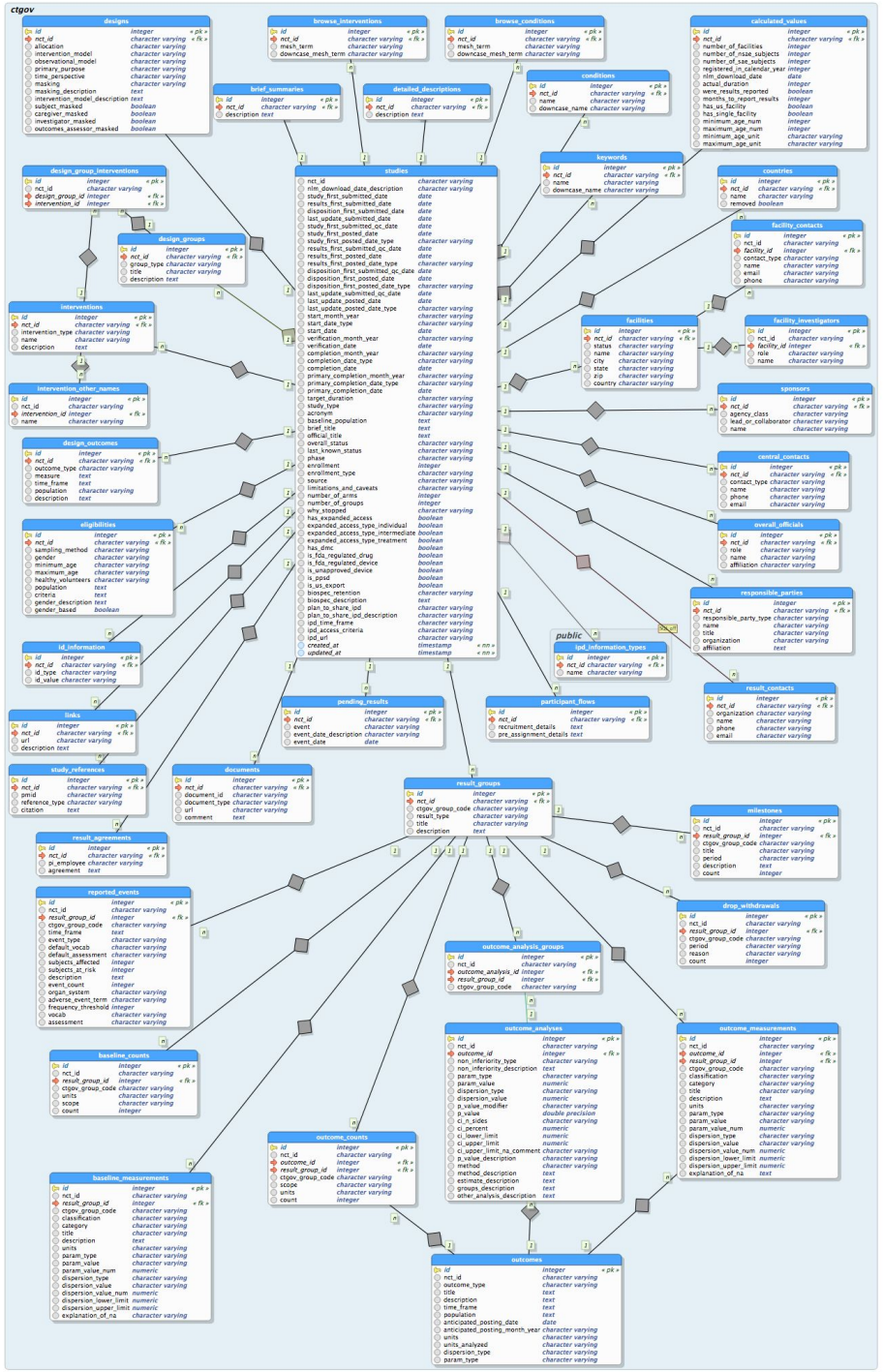
Diagnostic yield, postoperative complications, comorbidities and lenght of stay will be analyzed and compared following these procedures.

Glance on the Dataset:

- As per the dataset of 14/09/2018 available on clinicaltrials.gov ~250,000 Trials are available.
- We have focused on lung and heart diseases only. So working with 25536 trials.
- Dataset https://aact.ctticlinicaltrials.org/static/static_db_copies/daily/20180914_clinical_trials.zip

- Keywords:
- "heart", "coronary", "myocardial", "cardiovascular", "pulmonary", "stroke", "lung", "ventricular"

- Criteria used for clustering
- 1) Inclusion Criteria
- 2) Exclusion Criteria
- 3) Brief Summary



Improving trial representation

- As of now we have used bag of words as trail vector representation.
- Each word in corpus corresponds to a feature

Our aim is to obtain an efficient and compact vector representation of a trial. Currently we evaluate the quality of trial representation by the Clustering quality :

a) **Cluster internal validation metrics** - Gives better score to clusters having high intra-cluster similarity and low inter-cluster similarity. Measured by Silhouette coefficient, …

b) **Ground Truth from Trial database** - Percent of 1,2,3,5, 7,10 conditions or keywords overlap between a pair of trials lying within the same cluster(taken from SQL database)

Intuition :
Better the clustering quality, better will be the output of our Case-based search for trials

Now onwards, we try to improve our vector representation of trials.

Preprocessing steps: Indexing

id [PK] serial	nct_id character varying	sampling_method character varying	gender character varying	minimum_age character varying	maximum_age character varying
1	NCT03142841	''	All	N/A	N/A
2	NCT03142828	''	All	18 Years	N/A
4	NCT03142802	''	Male	18 Years	N/A
6	NCT03142776	''	All	18 Years	35 Years
7	NCT03142763	''	All	18 Years	30 Years
8	NCT03142750	''	All	31 Days	25 Years
10	NCT03142724	''	All	18 Years	55 Years
12	NCT03142698	''	All	18 Years	90 Years

- We partitioned the data into clusters on the bases of features like:
  gender , minAge, MaxAge

- We performed smoothing on data to get better clusters using some rules and assumptions

Before Smoothening:

```
(1, 10, 90) 9864
(2, 43, 81) 18
(2, 18, 52) 31
(3, 13, 44) 127
(1, 12, 84) 1812
(1, 10, 74) 1627
(2, 45, 78) 17
```

.
.
.
.

```
(2, 74, 83) 5
(2, 69, 83) 3
(2, 15, 480) 2
(3, 0, 365) 1
(4, 18, 74) 1
(2, 55, 55) 1
(3, 24, 70) 1
Size : 247
Smaller than 10 : 78
```

Clusters

After Smoothening:

```
(1, 10, 90) 10663
(2, 40, 80) 18
(2, 10, 50) 35
(3, 10, 40) 137
(1, 10, 80) 1862
(1, 10, 70) 1738
(2, 40, 70) 21
(1, 20, 80) 323
(2, 40, 60) 22
```

.
.

```
(2, 0, 20) 2
(2, 40, 40) 1
(3, 20, 20) 1
(3, 0, 0) 1
(2, 70, 70) 1
(2, 50, 50) 1
(2, 20, 70) 1
Size :  141
Smaller than 10 :  55
```

Clusters

Our approaches to improve clustering quality

Approach 1 : BitMap vector with cosine similarity

- For this we lemmatized , removed stop words and POS Tagged the words:

Numerical values, POS tags - CD, once, twice, thrice	Numeric
hz,kg,g,mmhg,ml,mg,mcg,cm,mm,doses,mmol,iu	units

- Current document representation : 3 separate vectors for exclusion(34252), inclusion(33952) and brief summary(52224) [Total length = 120428]
- We then removed all “numeric” terms to reduce size of dataset
- Vector size reduced by 1.2% by doing the same
- Bitmap vectors were compared using cosine similarity with threshold of 0.49 (empirically chosen)
- Formed individual set of clusters for each criteria
- Final Result is based on extent of overlap or intersection

Output found after intersection was too poor. (trials in same clusters were not related)

Approach 2 : TF-IDF with weighted Cosine similarity score

- Here, we define similarity score between a pair of trials “i” and “j” :
  Instead of creating separate set of clusters like previous approach, we will use this weighted similarity score to directly form our final clusters

$$SimScore(i, j) = \alpha * CosSim(incl\_vec_i, incl\_vec_j) + \beta * CosSim(excl\_vec_i, excl\_vec_j) + \gamma * CosSim(summ\_vec_i, summ\_vec_j)$$

  Taking $\alpha = \beta = \gamma = 1$ **and threshold value = 0.8**
  With such a threshold we found no of clusters to be too high(~5800) and each cluster had Trials(~4 trials a cluster) which were almost similar

- We are now working on finding more combination of α,β,γ and threshold to get better clusters.

ClinicalTrials.gov

Find Studies | About Studies | Submit Studies | Resources | About Site

Home > Search Results > Study Record Detail

Save this study

Trial record **1 of 1** for: NCT01977131

Previous Study | Return to List | Next Study

Safety Study of Autologous Bone Marrow Stromal Cells With Modification by Hepatocyte Growth Factor to Treat Silicosis

NIH U.S. National Library of Medicine
ClinicalTrials.gov

Find Studies | About Studies | Submit Studies | Resources | About Site

Home > Search Results > Study Record Detail

Save this study

Trial record **1 of 1** for: NCT01239862

Previous Study | Return to List | Next Study

Safety of Stem Cells Intrabronchial Instillation for Silicosis (SilicStemCell)

Approach 3 : Word embeddings

- We used a model like Doc2Vec to again represent the vectors of document according to brief summary, inclusion criteria and exclusion criteria.
- This time length of each vector was set to 300

```
data = nlp_clean(data)
it = LabeledLineSentence(data, docLabels)
model = gensim.models.Doc2Vec(size=300,)
```

- We considered brief summary vector as of now to find the cosine similarities between each 2 of random 100 Trials:

```
(0.1962253600358963, 6, 3372)
(0.19623124599456787, 6, 9646)
(0.19623728096485138, 6, 1755)
(0.1962418109178543, 6, 6649)
(0.19624201953411102, 5, 12017)
(0.19624298810958862, 3, 12070)
(0.19624726474285126, 6, 24893)
```

```
(0.7531717419624329, 4, 12660)
(0.7551890015602112, 4, 16769)
(0.759138822555542, 4, 17228)
(0.7812516093254089, 1, 21105)
(0.7848712205886841, 4, 10632)
(0.7887321710586548, 1, 11098)
(0.8182715177536011, 1, 21962)
```

- Now we will represent them as clusters using some empirically found threshold value.

Future works

- i) Work out some more ways to improve cluster quality
  - Dive deep into dataset and find manually which features weigh more to similarity
    - Interventions (description of drug/device used in trial)
    - Conditions list
    - etc
- ii) building a small search engine : given a nct id return its related nct ids