03_linking_trials_with_articles

March 16, 2025

[16]:
```
import pandas as pd
pubmed = pd.read_csv('data/all_pubmed.csv')
pubmed.head()
```

[16]:

	pubmed_id	title	keywords	journal	abstract	methods	results
0	40073121	Targeting the NPY/NPY1R signaling axis in muta…	NaN	Science advances	Pancreatic cancer (PC) is a highly metastatic …	NaN	NaN
1	40069621	The value of preoperative RDW for post-pancrea…	Pancreatic ductal adenocarcinoma, Post-pancrea…	BMC cancer	Pancreatic ductal adenocarcinoma (PDAC) is a h…	NaN	A total of 2268 patients were analyzed. We fou…
2	40069616	Protocol of the IMPACT study: randomized, mult…	Atezolizumab, Bevacizumab, Conversion, Hepatoc…	BMC cancer	Atezolizumab plus bevacizumab is recommended a…	NaN	NaN
3	40066089	Association between human leukocyte antigen E …	HLA-E, cancer, human leukocyte antigen, immuno…	Frontiers in oncology	Immunotherapy has gained momentum with the dis…	NaN	After screening 657 articles, 11 studies were …
4	40065459	Oncological and Survival Endpoints in Cancer C…	adverse events, cachexia, cancer, clinical tri…	Journal of cachexia, sarcopenia and muscle	In patients receiving anti-cancer treatment, c…	NaN	NaN

```
4  Fifty-seven trials were eligible, totalling 97…

                                        conclusions publication_date  \
0                                               NaN       2025-03-12
1  The preoperative RDW may be a useful marker fo…       2025-03-12
2                                               NaN       2025-03-12
3  This systematic review highlights that HLA-E e…       2025-03-11
4  In CC trials, oncological endpoints were mostl…       2025-03-11

                      category
0  Pubmed_Pancreatic_Cancer.csv
1  Pubmed_Pancreatic_Cancer.csv
2  Pubmed_Pancreatic_Cancer.csv
3  Pubmed_Pancreatic_Cancer.csv
4  Pubmed_Pancreatic_Cancer.csv
```

[17]:
```python
import pandas as pd
import re
import json

# Function to extract NCT IDs safely
def extract_nct_ids_from_abstract(abstract):
    """
    Extract all NCT IDs from the abstract using regex.
    Ensure that the input is a string before applying regex.
    """
    if not isinstance(abstract, str):  # Ensure it's a valid string
        return []
    return re.findall(r'NCT\d+', abstract)

# Function to link PubMed articles to Clinical Trials
def link_pubmed_to_trials(pubmed_df):
    """
    Link PubMed articles to Clinical Trials using NCT IDs extracted from the␣
 ↪abstracts.
    """
    data = []

    for index, row in pubmed_df.iterrows():
        pubmed_id = row['pubmed_id']
        abstract = row.get('abstract', '')  # Get abstract safely, default to␣
 ↪empty string
        nct_ids = extract_nct_ids_from_abstract(abstract)

        if nct_ids:  # Only add if there are valid NCT IDs
            data.append({"PubMed_ID": pubmed_id, "NCT_IDs": nct_ids})
```

```
    return data

# Extracting and saving to JSON
linked_data = link_pubmed_to_trials(pubmed)

# Saving results to JSON
output_file = "data/linked_pubmed_nct_ids.json"
with open(output_file, 'w') as json_file:
    json.dump(linked_data, json_file, indent=4)

print(f"Data saved in {output_file}")
```

```
Data saved in data/linked_pubmed_nct_ids.json
```

[13]:
```
import pandas as pd
import json

# pd.read_csv('data/all_diseas_processed.csv')
all_disease = pd.read_csv('data/CT_all_common_disease_processed.csv')
print(all_disease.head())

pubmed_links = json.load(open('linked_pubmed_nct_ids.json'))
print(pubmed_links[:5])




# Create a mapping of NCT ID to PubMed ID
nct_to_pubmed = {}

for entry in pubmed_links:
    pubmed_id = str(entry["PubMed_ID"])  # Ensure string format
    for nct_id in entry["NCT_IDs"]:
        nct_to_pubmed[nct_id] = pubmed_id

# Sample all_disease DataFrame
# all_disease = pd.DataFrame({
#     "NCT ID": ["NCT06088706", "NCT02871856", "NCT12345678", "NCT05622630"]
# })

# Map NCT ID to PubMed ID
all_disease["Associated Article ID"] = all_disease["NCT ID"].map(nct_to_pubmed).
 ↪fillna("")

# Create binary column indicating association
all_disease["Associated Article?"] = all_disease["Associated Article ID"].
 ↪apply(lambda x: "YES" if x else "NO")
```

```
# Display result
print(all_disease)

print("Count of all common diseases:", len(all_disease))
print("Number of trials with associated articles:", all_disease['Associated␣
 ↪Article?'].value_counts()['YES'])
print("Number of trials not having any associated articles:",␣
 ↪all_disease['Associated Article?'].value_counts()['NO'])
```

```
       NCT ID  Acronym         Overall Status  Start Date  \
0  NCT03116126    NorAD  ACTIVE_NOT_RECRUITING  2019-01-04
1  NCT04137926  Unknown                UNKNOWN  2020-03-01
2  NCT02537626  Unknown              COMPLETED  2018-03-15
3  NCT05531526  Unknown              RECRUITING  2022-12-23
4  NCT00297362  Unknown              COMPLETED     2004-06

            Conditions                       Interventions  \
0    Alzheimer Disease                Guanfacine, Placebo
1  Alzheimer's Disease                    MicRNAs battery
2  Alzheimer's Disease  Erchonia ALS Laser, Placebo Laser
3    Alzheimer Disease                    AR1001, Placebo
4    Alzheimer Disease           Galantamine hydrobromide

                                      Locations Primary Completion Date  \
0                          London - United Kingdom              2024-08-05
1                                Shanghai - China              2022-08-30
2            Zapopan - Mexico, La Plazuela - Mexico              2020-06-15
3  Phoenix - United States, Scottsdale - United S…                2025-12
4                             No locations listed            Unknown Date

  Study First Post Date Last Update Post Date      Study Type         Phases  \
0            2017-04-14            2024-05-13  INTERVENTIONAL         PHASE3
1            2019-10-24            2021-09-23  INTERVENTIONAL            NaN
2            2015-09-01            2021-05-25  INTERVENTIONAL            NaN
3            2022-09-08            2025-02-07  INTERVENTIONAL         PHASE3
4            2006-02-28            2012-03-26   OBSERVATIONAL  Not Available

                                   Sponsor Sponsor Type         Disease
0                     Imperial College London        OTHER  CT_alzheimer's
1                Shanghai Mental Health Center        OTHER  CT_alzheimer's
2                        Erchonia Corporation     INDUSTRY  CT_alzheimer's
3                            AriBio Co., Ltd.     INDUSTRY  CT_alzheimer's
4  Janssen Cilag Pharmaceutica S.A.C.I., Greece     INDUSTRY  CT_alzheimer's
[{'PubMed_ID': 40060302, 'NCT_IDs': ['NCT03270917']}, {'PubMed_ID': 40055837,
'NCT_IDs': ['NCT03876561']}, {'PubMed_ID': 39846984, 'NCT_IDs':
['NCT03331562']}, {'PubMed_ID': 39834129, 'NCT_IDs': ['NCT02795858']},
```

```
{'PubMed_ID': 39833722, 'NCT_IDs': ['NCT04241276']}]
            NCT ID    Acronym         Overall Status  Start Date  \
0      NCT03116126      NorAD  ACTIVE_NOT_RECRUITING  2019-01-04
1      NCT04137926    Unknown                UNKNOWN  2020-03-01
2      NCT02537626    Unknown              COMPLETED  2018-03-15
3      NCT05531526    Unknown              RECRUITING  2022-12-23
4      NCT00297362    Unknown              COMPLETED     2004-06
…            …         …                      …          …
14113  NCT05842954     KALUMA              RECRUITING  2024-03-07
14114  NCT06232954  Mossie-GO              RECRUITING  2024-06-05
14115  NCT01422954    Unknown              COMPLETED     2012-01
14116  NCT00113854    Unknown                UNKNOWN     2004-10
14117  NCT01916954      ALN5P              COMPLETED     2013-07

                                      Conditions  \
0                                Alzheimer Disease
1                              Alzheimer's Disease
2                              Alzheimer's Disease
3                                Alzheimer Disease
4                                Alzheimer Disease
…                                              …
14113  Uncomplicated Plasmodium Falciparum Malaria
14114              Malaria, Mosquito-Borne Disease
14115               Malaria, Plasmodium Falciparum
14116                              Cerebral Malaria
14117                                       Malaria

                                           Interventions  \
0                                    Guanfacine, Placebo
1                                        MicRNAs battery
2                       Erchonia ALS Laser, Placebo Laser
3                                    AR1001, Placebo
4                               Galantamine hydrobromide
…                                                    …
14113                                KLU156, Coartem
14114  Mossie-Go containing treated transfluthrin dis…
14115  Chloroquine prophylaxis, Mefloquine prophylaxi…
14116                                          Mannitol
14117  3-day artemether-lumefantrine, 5-day artemethe…

                                                Locations  \
0                                London - United Kingdom
1                                       Shanghai - China
2                   Zapopan - Mexico, La Plazuela - Mexico
3      Phoenix - United States, Scottsdale - United S…
4                                   No locations listed
…                                                      …
14113  Bobo Dioulasso - Burkina Faso, Nanoro - Burkin…
```

```
14114                                  Jinja - Uganda
14115                               Leiden - Netherlands
14116                                  Kampala - Uganda
14117                                  Kinshasa - Congo

      Primary Completion Date Study First Post Date Last Update Post Date  \
0                  2024-08-05            2017-04-14            2024-05-13
1                  2022-08-30            2019-10-24            2021-09-23
2                  2020-06-15            2015-09-01            2021-05-25
3                     2025-12            2022-09-08            2025-02-07
4                Unknown Date            2006-02-28            2012-03-26
…                         …                     …                     …
14113              2025-07-07            2023-05-06            2024-12-31
14114                 2025-09            2024-01-31            2024-06-21
14115                 2012-12            2011-08-25            2013-04-29
14116            Unknown Date            2005-06-13            2005-06-24
14117                 2014-03            2013-08-06            2014-03-26

          Study Type         Phases  \
0      INTERVENTIONAL         PHASE3
1      INTERVENTIONAL            NaN
2      INTERVENTIONAL            NaN
3      INTERVENTIONAL         PHASE3
4       OBSERVATIONAL  Not Available
…                 …              …
14113  INTERVENTIONAL         PHASE3
14114  INTERVENTIONAL            NaN
14115  INTERVENTIONAL            NaN
14116  INTERVENTIONAL         PHASE3
14117  INTERVENTIONAL         PHASE3

                                          Sponsor Sponsor Type  \
0                         Imperial College London        OTHER
1                  Shanghai Mental Health Center        OTHER
2                           Erchonia Corporation     INDUSTRY
3                               AriBio Co., Ltd.     INDUSTRY
4      Janssen Cilag Pharmaceutica S.A.C.I., Greece     INDUSTRY
…                                              …            …
14113                    Novartis Pharmaceuticals     INDUSTRY
14114                        Africa Power Limited     INDUSTRY
14115          Radboud University Medical Center        OTHER
14116                        Makerere University        OTHER
14117                        University of Oxford        OTHER

             Disease Associated Article ID Associated Article?
0      CT_alzheimer's                                       NO
1      CT_alzheimer's                                       NO
2      CT_alzheimer's                                       NO
```

```
3      CT_alzheimer's                                    NO
4      CT_alzheimer's                                    NO
…                 …                    …                 …
14113      CT_malaria                                    NO
14114      CT_malaria                                    NO
14115      CT_malaria             25396417               YES
14116      CT_malaria                                    NO
14117      CT_malaria                                    NO

[14118 rows x 17 columns]
Count of all common diseases: 14118
Number of trials with associated articles: 1137
Number of trials not having any associated articles: 12981
```

[14]:
```python
import pandas as pd
import json

# pd.read_csv('data/all_diseas_processed.csv')
all_disease = pd.read_csv('data/CT_all_rare_disease_processed.csv')
print(all_disease.head())

pubmed_links = json.load(open('linked_pubmed_nct_ids.json'))
print(pubmed_links[:5])




# Create a mapping of NCT ID to PubMed ID
nct_to_pubmed = {}

for entry in pubmed_links:
    pubmed_id = str(entry["PubMed_ID"])  # Ensure string format
    for nct_id in entry["NCT_IDs"]:
        nct_to_pubmed[nct_id] = pubmed_id

# Sample all_disease DataFrame
# all_disease = pd.DataFrame({
#     "NCT ID": ["NCT06088706", "NCT02871856", "NCT12345678", "NCT05622630"]
# })

# Map NCT ID to PubMed ID
all_disease["Associated Article ID"] = all_disease["NCT ID"].map(nct_to_pubmed).
 ↪fillna("")

# Create binary column indicating association
all_disease["Associated Article?"] = all_disease["Associated Article ID"].
 ↪apply(lambda x: "YES" if x else "NO")
```

```
# Display result
print(all_disease)

print("Count of all rare diseases:", len(all_disease))
print("Number of trials with associated articles:", all_disease['Associated␣
 ↪Article?'].value_counts()['YES'])
print("Number of trials not having any associated articles:",␣
 ↪all_disease['Associated Article?'].value_counts()['NO'])
```

```
        NCT ID   Acronym Overall Status    Start Date  \
0  NCT00005926   Unknown      COMPLETED       2000-06
1  NCT01094626   Unknown      WITHDRAWN       2010-04
2  NCT00253526   Unknown      WITHDRAWN  Unknown Date
3  NCT00003426   Unknown      COMPLETED       1998-04
4  NCT03469726  DIA-PANC        UNKNOWN    2017-12-22

                                        Conditions  \
0            Pancreatic Cancer, Pancreatic Neoplasm
1  Pancreatic Cancer, Intraductal Papillary Mucin…
2  Adenocarcinoma of the Pancreas, Recurrent Panc…
3                                Pancreatic Cancer
4                              Pancreatic Neoplasms

                                     Interventions  \
0         Gemcitabine, Herceptin, Radiation therapy
1                                          Secretin
2  bevacizumab, gemcitabine hydrochloride, adjuva…
3      gemcitabine hydrochloride, radiation therapy
4          Contrast-enhanced Diffusion-weighted MRI

                                       Locations Primary Completion Date  \
0                         Bethesda - United States            Unknown Date
1                              No locations listed                 2013-03
2                              No locations listed            Unknown Date
3                         New York - United States                 2002-03
4  Athens - Greece, Nijmegen - Netherlands, Den B…              2022-01-01

  Study First Post Date Last Update Post Date      Study Type  Phases  \
0            2002-12-10            2008-03-04  INTERVENTIONAL  PHASE2
1            2010-03-29            2016-06-10  INTERVENTIONAL     NaN
2            2005-11-15            2015-04-28  INTERVENTIONAL  PHASE2
3            2003-12-02            2013-06-21  INTERVENTIONAL  PHASE1
4            2018-03-19            2021-09-21  INTERVENTIONAL     NaN

                                Sponsor Sponsor Type              Disease
0        National Cancer Institute (NCI)          NIH  CT_pancreatic_cancer
```

```
1                         Elizabeth Hecht        OTHER  CT_pancreatic_cancer
2         National Cancer Institute (NCI)          NIH  CT_pancreatic_cancer
3  Memorial Sloan Kettering Cancer Center        OTHER  CT_pancreatic_cancer
4       Radboud University Medical Center        OTHER  CT_pancreatic_cancer
[{'PubMed_ID': 40060302, 'NCT_IDs': ['NCT03270917']}, {'PubMed_ID': 40055837,
'NCT_IDs': ['NCT03876561']}, {'PubMed_ID': 39846984, 'NCT_IDs':
['NCT03331562']}, {'PubMed_ID': 39834129, 'NCT_IDs': ['NCT02795858']},
{'PubMed_ID': 39833722, 'NCT_IDs': ['NCT04241276']}]
           NCT ID    Acronym Overall Status    Start Date  \
0     NCT00005926    Unknown      COMPLETED       2000-06
1     NCT01094626    Unknown      WITHDRAWN       2010-04
2     NCT00253526    Unknown      WITHDRAWN  Unknown Date
3     NCT00003426    Unknown      COMPLETED       1998-04
4     NCT03469726   DIA-PANC        UNKNOWN    2017-12-22
…           …        …              …            …
3643  NCT00312247    Unknown      COMPLETED       2006-04
3644  NCT02020954  Becker-HS        UNKNOWN       2013-01
3645  NCT05833633    Unknown     RECRUITING    2022-03-18
3646  NCT04740554    Unknown      COMPLETED    2013-03-01
3647  NCT01350154    Unknown      COMPLETED       2011-11

                                          Conditions  \
0               Pancreatic Cancer, Pancreatic Neoplasm
1     Pancreatic Cancer, Intraductal Papillary Mucin…
2     Adenocarcinoma of the Pancreas, Recurrent Panc…
3                                   Pancreatic Cancer
4                                Pancreatic Neoplasms
…                                                  …
3643                       Duchenne Muscular Dystrophy
3644  Dilated Cardiomyopathy, Lef Ventricular Dysfun…
3645                      Muscular Dystrophy, Duchenne
3646                       Duchenne Muscular Dystrophy
3647                         Becker Muscular Dystrophy

                                       Interventions  \
0            Gemcitabine, Herceptin, Radiation therapy
1                                             Secretin
2     bevacizumab, gemcitabine hydrochloride, adjuva…
3          gemcitabine hydrochloride, radiation therapy
4              Contrast-enhanced Diffusion-weighted MRI
…                                                  …
3643                         No interventions listed
3644  ECG, echocardiography, cardiac MRI, sera bioma…
3645           MUSCLE MAGNETIC RESONANCE IMAGING (MRI)
3646  Duchenne Muscular Dystrophy group with Deflaza…
3647                                Sildenafil, Placebo

                                             Locations  \
```

```
0                              Bethesda - United States
1                                   No locations listed
2                                   No locations listed
3                              New York - United States
4     Athens - Greece, Nijmegen - Netherlands, Den B…
…                                                   …
3643  Los Angeles - United States, Sacramento - Unit…
3644                                      Paris - France
3645                                        Rome - Italy
3646                                   No locations listed
3647                                  Copenhagen - Denmark

     Primary Completion Date Study First Post Date Last Update Post Date  \
0               Unknown Date            2002-12-10            2008-03-04
1                    2013-03            2010-03-29            2016-06-10
2               Unknown Date            2005-11-15            2015-04-28
3                    2002-03            2003-12-02            2013-06-21
4                  2022-01-01            2018-03-19            2021-09-21
…                         …                     …                     …
3643                 2014-12            2006-04-07            2015-05-19
3644                 2017-06            2013-12-25            2015-03-24
3645               2023-04-17            2023-04-27            2023-05-15
3646               2014-09-01            2021-02-05            2021-02-05
3647                 2013-04            2011-05-09            2013-04-10

          Study Type         Phases  \
0     INTERVENTIONAL         PHASE2
1     INTERVENTIONAL            NaN
2     INTERVENTIONAL         PHASE2
3     INTERVENTIONAL         PHASE1
4     INTERVENTIONAL            NaN
…                 …              …
3643   OBSERVATIONAL  Not Available
3644   OBSERVATIONAL  Not Available
3645   OBSERVATIONAL  Not Available
3646   OBSERVATIONAL  Not Available
3647  INTERVENTIONAL         PHASE2

                                                 Sponsor Sponsor Type  \
0                       National Cancer Institute (NCI)          NIH
1                                     Elizabeth Hecht        OTHER
2                       National Cancer Institute (NCI)          NIH
3                  Memorial Sloan Kettering Cancer Center        OTHER
4                       Radboud University Medical Center        OTHER
…                                                     …          …
3643                    Shriners Hospitals for Children        OTHER
3644                                       Karim WAHBI        OTHER
3645  Fondazione Policlinico Universitario Agostino …        OTHER
```

```
3646                          University of Sao Paulo        OTHER
3647                          Rigshospitalet, Denmark        OTHER

                            Disease Associated Article ID Associated Article?
0               CT_pancreatic_cancer                                       NO
1               CT_pancreatic_cancer                                       NO
2               CT_pancreatic_cancer                                       NO
3               CT_pancreatic_cancer                                       NO
4               CT_pancreatic_cancer                                       NO
…                                  …                  …                   …
3643  CT_duchenne_muscular_dystrophy                                       NO
3644  CT_duchenne_muscular_dystrophy                                       NO
3645  CT_duchenne_muscular_dystrophy                                       NO
3646  CT_duchenne_muscular_dystrophy                                       NO
3647  CT_duchenne_muscular_dystrophy                                       NO

[3648 rows x 17 columns]
Count of all rare diseases: 3648
Number of trials with associated articles: 422
Number of trials not having any associated articles: 3226
```