University of BRISTOL

Elizabeth Blackwell Insititute for Health Research

Data acquisition and pilot study on BioRxiv and MedRxiv full text data to facilitate comprehensive data mining on biomedical literature

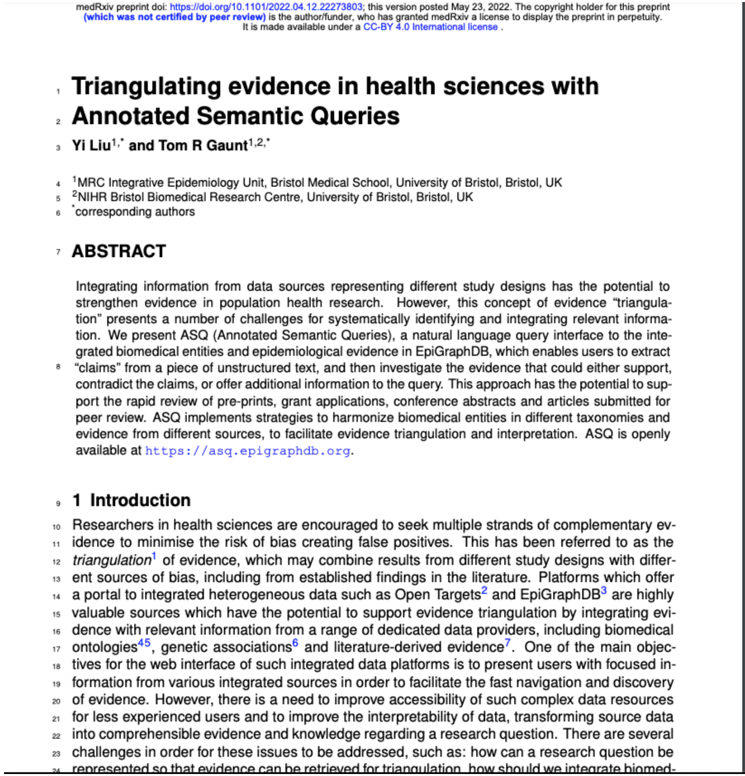
A project funded by
Elizabeth Blackwell Institute
Rapid Research Call

Yi Liu

IEU Programme 3 meeting
15 August 2023

Background

- The proposed project idea originated from our earlier work ASQ (Liu & Gaunt, 2022)
- We scraped and analysed MedRxiv abstracts from 2020-01-01 to 2021-12-31
- Limits:
  - Web scraping 24 months of *just* abstracts was time consuming
  - MedRxiv API only keeps one version of metadata
- We have known BioRxiv / MedRxiv kept full text data for text mining



Triangulating evidence in health sciences with Annotated Semantic Queries

Yi Liu[1,*] and Tom R Gaunt[1,2,*]

[1]MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK
[2]NIHR Bristol Biomedical Research Centre, University of Bristol, Bristol, UK
[*]corresponding authors

ABSTRACT

Integrating information from data sources representing different study designs has the potential to strengthen evidence in population health research. However, this concept of evidence "triangulation" presents a number of challenges for systematically identifying and integrating relevant information. We present ASQ (Annotated Semantic Queries), a natural language query interface to the integrated biomedical entities and epidemiological evidence in EpiGraphDB, which enables users to extract "claims" from a piece of unstructured text, and then investigate the evidence that could either support, contradict the claims, or offer additional information to the query. This approach has the potential to support the rapid review of pre-prints, grant applications, conference abstracts and articles submitted for peer review. ASQ implements strategies to harmonize biomedical entities in different taxonomies and evidence from different sources, to facilitate evidence triangulation and interpretation. ASQ is openly available at https://asq.epigraphdb.org.

1 Introduction

Researchers in health sciences are encouraged to seek multiple strands of complementary evidence to minimise the risk of bias creating false positives. This has been referred to as the *triangulation*[1] of evidence, which may combine results from different study designs with different sources of bias, including from established findings in the literature. Platforms which offer a portal to integrated heterogeneous data such as Open Targets[2] and EpiGraphDB[3] are highly valuable sources which have the potential to support evidence triangulation by integrating evidence with relevant information from a range of dedicated data providers, including biomedical ontologies[45], genetic associations[6] and literature-derived evidence[7]. One of the main objectives for the web interface of such integrated data platforms is to present users with focused information from various integrated sources in order to facilitate the fast navigation and discovery of evidence. However, there is a need to improve accessibility of such complex data resources for less experienced users and to improve the interpretability of data, transforming source data into comprehensible evidence and knowledge regarding a research question. There are several challenges in order for these issues to be addressed, such as: how can a research question be represented so that evidence can be retrieved for triangulation, how should we integrate biomed-

About the project

- This is a seedcorn funding project funded by the Elizabeth Blackwell Institute 2023 Rapid Research Call to us (**Yi Liu**, Tom Gaunt)
- We proposed to acquire the full text data of BioRxiv and MedRxiv preprints and conduct pilot studies on the acquired data
- Data and results from this project will lead to our next stage projects involving cross-Faculty collaborations (in progress)
- Code on processing and analysis https://github.com/mrcieu/biorxiv-medrxiv-tdm

Timeline (2023-06-05 -- 2023-07-31)

University of BRISTOL
Elizabeth Blackwell Insititute for Health Research

- 05-31 Award confirmed; 06-08 Budget code generated;
- 06-09 -- 06-23 Tried to sort out payment mechanisms
    - Finance; IT; Finance; Procurement; AWS; Procurement
- 06-26 -- 06-29 AWS access setup;
  Data transfer from S3 to epi-franklin
- 07-01 -- 07-31 Exploratory analysis
- 07-01 -- 07-17 GDPR compliance check;
- 07-19 -- 07-27 Contact ACRC on purchasing RDSF;
- 07-02 -- 07-27 Setting up MyERP things (requisition, purchase order, etc.) to get the invoice paid
- 08-02 -- ... Chasing Finance to get the invoice paid

Why didn't I ...

- Set up AWS things sooner?
    - Need to appropriately set up University procurement / payment process
    - Need experiments on costs with small batches
- Contact ACRC sooner?
    - GDPR compliance check on individual identifiable information

Lessons

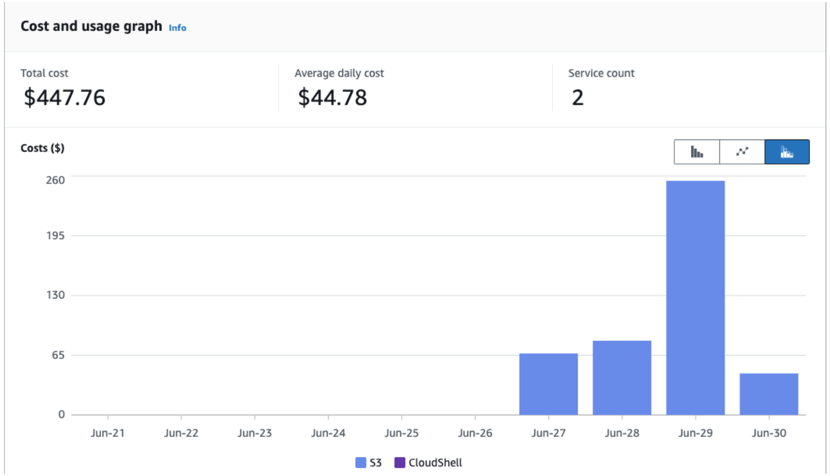
- Should have asked around for prior experiences & lessons more

Text/data mining BioRxiv / MedRxiv

- Via API: 1) List of submitted preprints based on a query time interval 2) metadata on individual preprint
- Via web scraping, based on the doi known from metadata
- Via full text data archives
  - Hosted on as AWS S3 Requester Pays buckets -- Requester pays for the costs associated with the data transfer
  - BioRxiv s3://biorxiv-src-monthly; MedRxiv s3://medrxiv-src-monthly

Amazon AWS

- Root user: University invoice account
    - MFA by a Google Authenticator ☺
- IAM user: sub-user specific for S3 access
- Could reuse for future projects
- Originally budgeted for
    - Transferring and storing on our buckets
    - Trasnferring out of AWS
    - 15T per month, two months



Cost and usage graph Info

Total cost $447.76

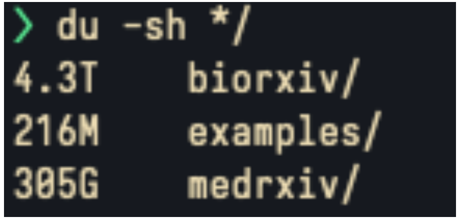
Average daily cost $44.78

Service count 2

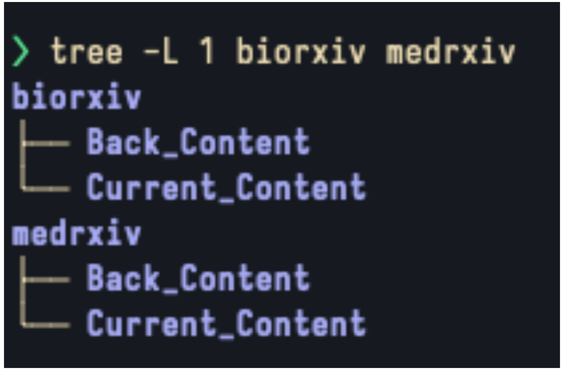
- Traffics on 4.5T + experiment batches of data access

University of BRISTOL
Elizabeth Blackwell Insititute for Health Research

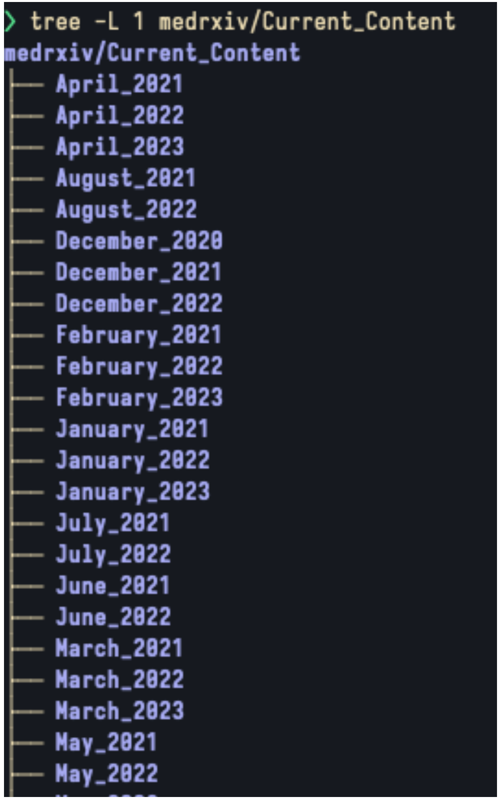
Acquired Dataset



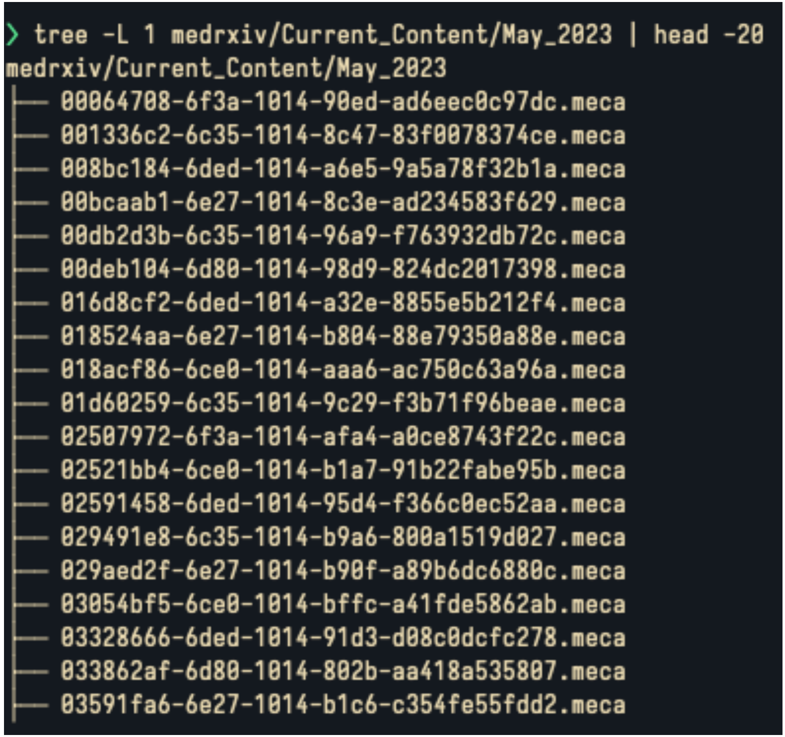
```
❯ du -sh */
4.3T    biorxiv/
216M    examples/
305G    medrxiv/
```



```
❯ tree -L 1 biorxiv medrxiv
biorxiv
├── Back_Content
└── Current_Content
medrxiv
├── Back_Content
└── Current_Content
```

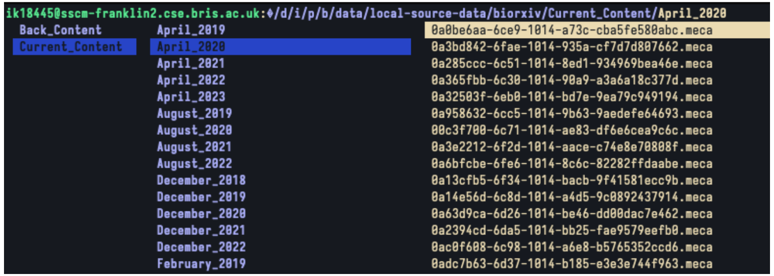


```
❯ tree -L 1 medrxiv/Current_Content
medrxiv/Current_Content
├── April_2021
├── April_2022
├── April_2023
├── August_2021
├── August_2022
├── December_2020
├── December_2021
├── December_2022
├── February_2021
├── February_2022
├── February_2023
├── January_2021
├── January_2022
├── January_2023
├── July_2021
├── July_2022
├── June_2021
├── June_2022
├── March_2021
├── March_2022
├── March_2023
├── May_2021
├── May_2022
```

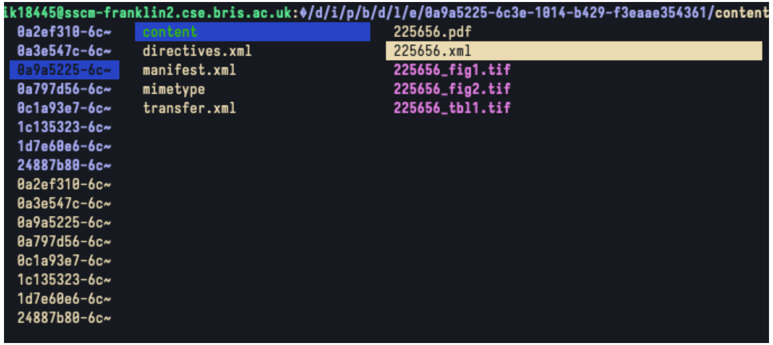


```
❯ tree -L 1 medrxiv/Current_Content/May_2023 | head -20
medrxiv/Current_Content/May_2023
├── 00064708-6f3a-1014-90ed-ad6eec0c97dc.meca
├── 001336c2-6c35-1014-8c47-83f0078374ce.meca
├── 008bc184-6ded-1014-a6e5-9a5a78f32b1a.meca
├── 00bcaab1-6e27-1014-8c3e-ad234583f629.meca
├── 00db2d3b-6c35-1014-96a9-f763932db72c.meca
├── 00deb104-6d80-1014-98d9-824dc2017398.meca
├── 016d8cf2-6ded-1014-a32e-8855e5b212f4.meca
├── 018524aa-6e27-1014-b804-88e79350a88e.meca
├── 018acf86-6ce0-1014-aaa6-ac750c63a96a.meca
├── 01d60259-6c35-1014-9c29-f3b71f96beae.meca
├── 02507972-6f3a-1014-afa4-a0ce8743f22c.meca
├── 02521bb4-6ce0-1014-b1a7-91b22fabe95b.meca
├── 02591458-6ded-1014-95d4-f366c0ec52aa.meca
├── 029491e8-6c35-1014-b9a6-800a1519d027.meca
├── 029aed2f-6e27-1014-b90f-a89b6dc6880c.meca
├── 03054bf5-6ce0-1014-bffc-a41fde5862ab.meca
├── 03328666-6ded-1014-91d3-d08c0dcfc278.meca
├── 033862af-6d80-1014-802b-aa418a535807.meca
├── 03591fa6-6e27-1014-b1c6-c354fe55fdd2.meca
```

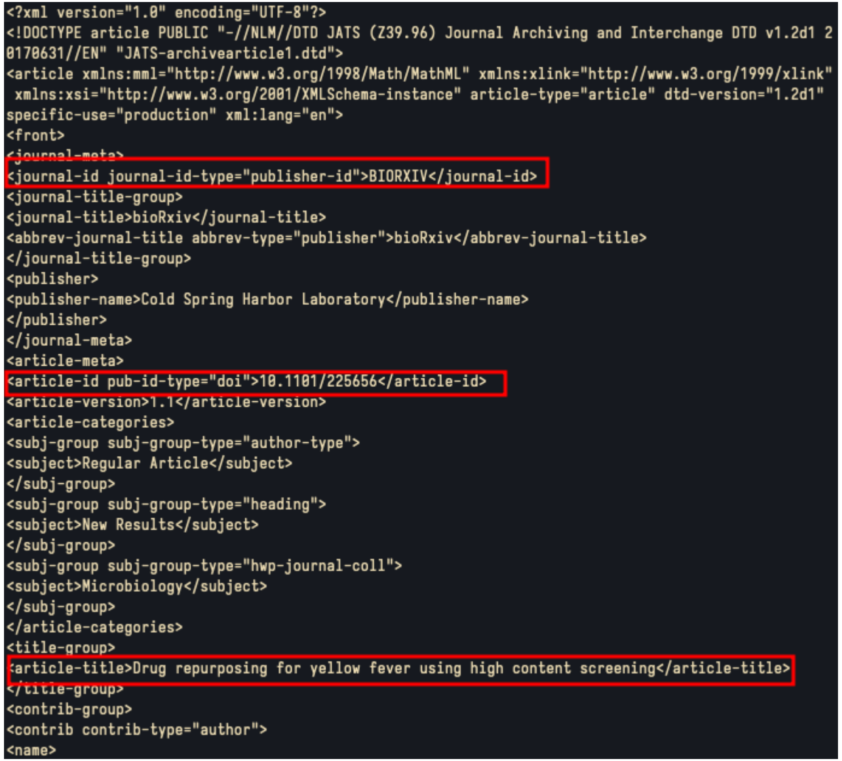
A data archive



```
ik18445@sscm-franklin2.cse.bris.ac.uk:⧫/d/i/p/b/data/local-source-data/biorxiv/Current_Content/April_2020
 Back_Content        April_2019                          0a0be6aa-6ce9-1014-a73c-cba5fe580abc.meca
 Current_Content     April_2020                          0a3bd842-6fae-1014-935a-cf7d7d807662.meca
                     April_2021                          0a285ccc-6c51-1014-8ed1-934969bea46e.meca
                     April_2022                          0a365fbb-6c30-1014-90a9-a3a6a18c377d.meca
                     April_2023                          0a32503f-6eb0-1014-bd7e-9ea79c949194.meca
                     August_2019                         0a958632-6cc5-1014-9b63-9aedefe64693.meca
                     August_2020                         00c3f700-6c71-1014-ae83-df6e6cea9c6c.meca
                     August_2021                         0a3e2212-6f2d-1014-aace-c74e8e70808f.meca
                     August_2022                         0a6bfcbe-6fe6-1014-8c6c-82282ffdaabe.meca
                     December_2018                       0a13cfb5-6f34-1014-bacb-9f41581ecc9b.meca
                     December_2019                       0a14e56d-6c8d-1014-a4d5-9c0892437914.meca
                     December_2020                       0a63d9ca-6d26-1014-be46-dd00dac7e462.meca
                     December_2021                       0a2394cd-6da5-1014-bb25-fae9579eefb0.meca
                     December_2022                       0ac0f608-6c98-1014-a6e8-b5765352ccd6.meca
                     February_2019                       0adc7b63-6d37-1014-b185-e3e3e744f963.meca
```

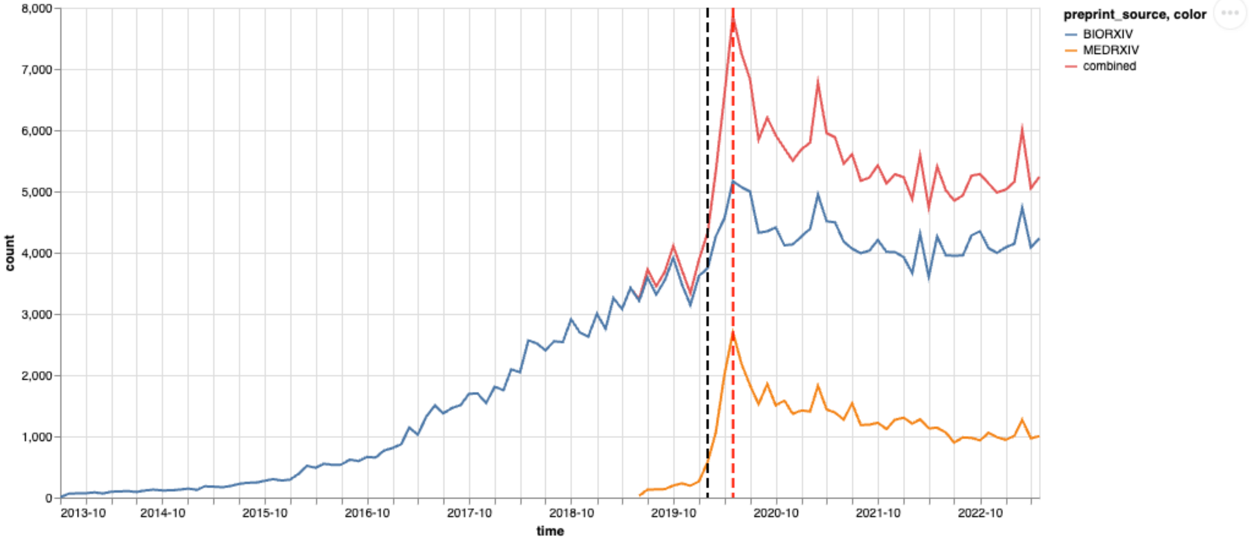


```
ik18445@sscm-franklin2.cse.bris.ac.uk:⧫/d/i/p/b/d/l/e/0a9a5225-6c3e-1014-b429-f3eaae354361/content
 0a2ef310-6c~   content                   225656.pdf
 0a3e547c-6c~   directives.xml            225656.xml
 0a9a5225-6c~   manifest.xml              225656_fig1.tif
 0a797d56-6c~   mimetype                  225656_fig2.tif
 0c1a93e7-6c~   transfer.xml              225656_tbl1.tif
 1c135323-6c~
 1d7e60e6-6c~
 24887b80-6c~
 0a2ef310-6c~
 0a3e547c-6c~
 0a9a5225-6c~
 0a797d56-6c~
 0c1a93e7-6c~
 1c135323-6c~
 1d7e60e6-6c~
 24887b80-6c~
```



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE article PUBLIC "-//NLM//DTD JATS (Z39.96) Journal Archiving and Interchange DTD v1.2d1 2
0170631//EN" "JATS-archivearticle1.dtd">
<article xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:xlink="http://www.w3.org/1999/xlink"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" article-type="article" dtd-version="1.2d1"
specific-use="production" xml:lang="en">
<front>
<journal-meta>
<journal-id journal-id-type="publisher-id">BIORXIV</journal-id>
<journal-title-group>
<journal-title>bioRxiv</journal-title>
<abbrev-journal-title abbrev-type="publisher">bioRxiv</abbrev-journal-title>
</journal-title-group>
<publisher>
<publisher-name>Cold Spring Harbor Laboratory</publisher-name>
</publisher>
</journal-meta>
<article-meta>
<article-id pub-id-type="doi">10.1101/225656</article-id>
<article-version>1.1</article-version>
<article-categories>
<subj-group subj-group-type="author-type">
<subject>Regular Article</subject>
</subj-group>
<subj-group subj-group-type="heading">
<subject>New Results</subject>
</subj-group>
<subj-group subj-group-type="hwp-journal-coll">
<subject>Microbiology</subject>
</subj-group>
</article-categories>
<title-group>
<article-title>Drug repurposing for yellow fever using high content screening</article-title>
</title-group>
<contrib-group>
<contrib contrib-type="author">
<name>
```

EDA: Volume

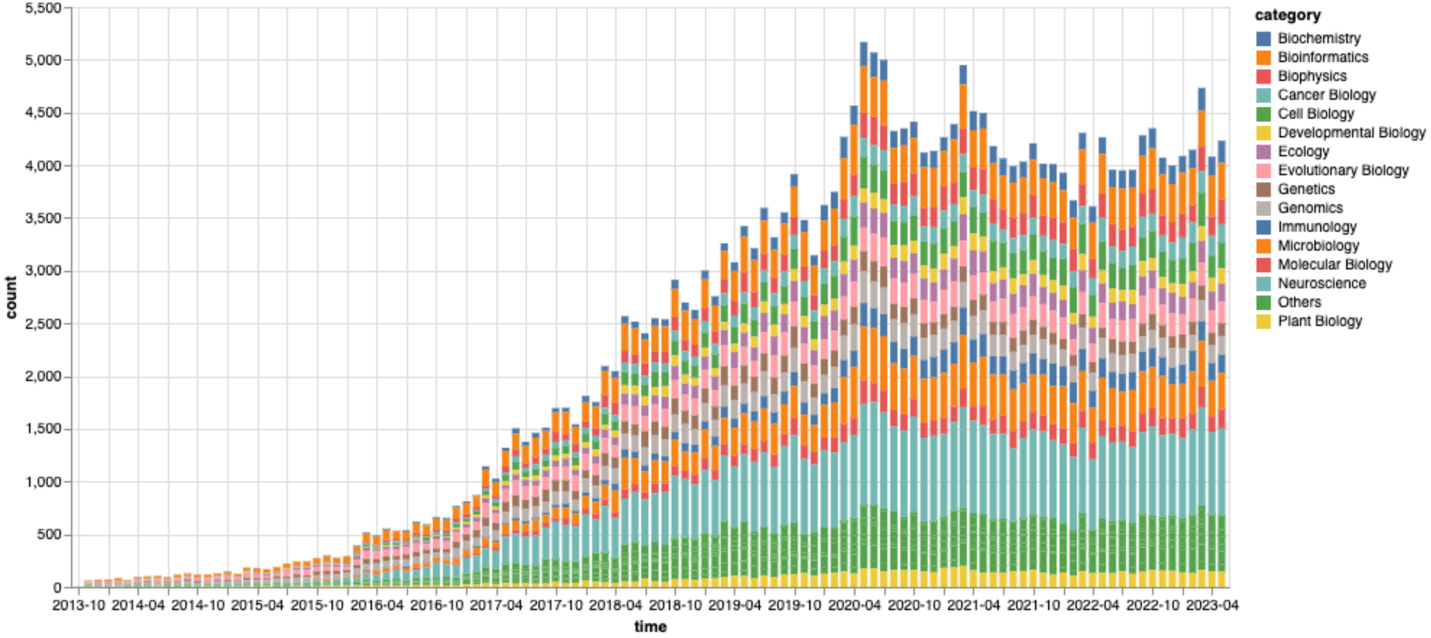


- The separation occurred in mid 2019
- Black line: 2020-02-01
- Red line: 2020-05-01

University of BRISTOL

Elizabeth Blackwell Insititute
for Health Research

EDA: Categories, BioRxiv

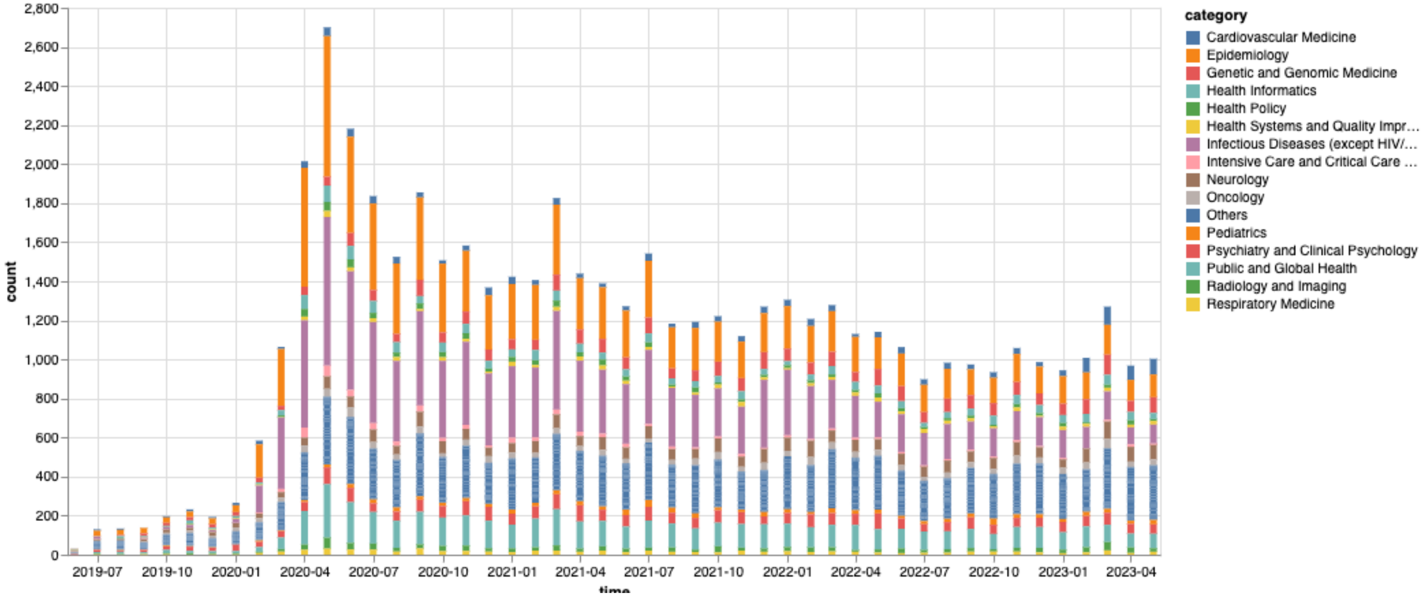


Top15 categories
+ Others

Most populous categories:

- Bioinformatics
- Microbiology
- Neuroscience

EDA: Categories, MedRxiv

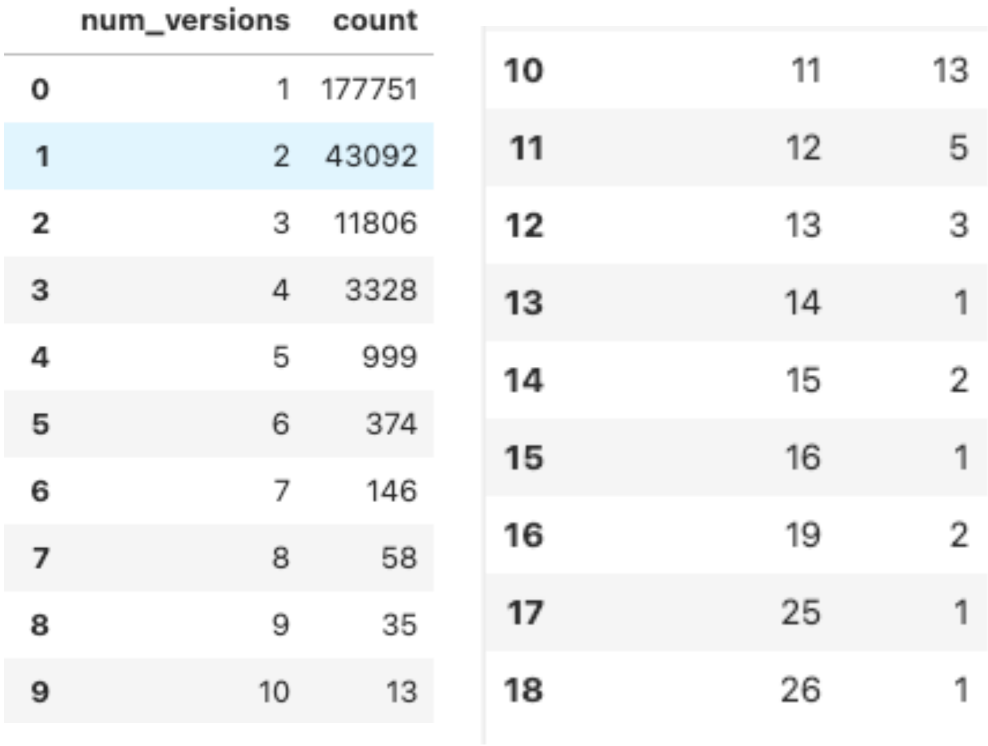


Top 15 categories + Others

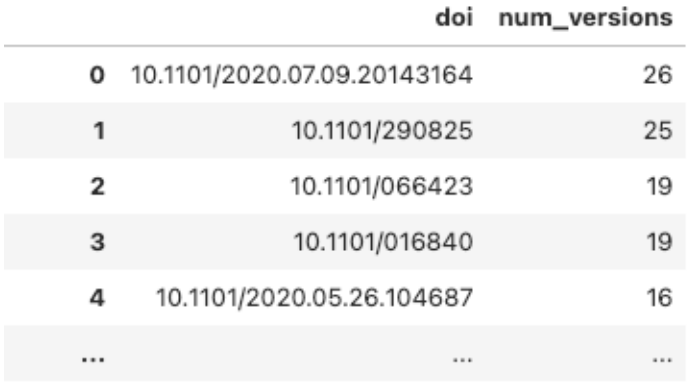
Most populous categories:

- Epidemiology
- Infectious Diseases (execpt HIV/AIDS)
- Public and Global Health

EDA: revisions



	num_versions	count
0	1	177751
1	2	43092
2	3	11806
3	4	3328
4	5	999
5	6	374
6	7	146
7	8	58
8	9	35
9	10	13
10	11	13
11	12	5
12	13	3
13	14	1
14	15	2
15	16	1
16	19	2
17	25	1
18	26	1

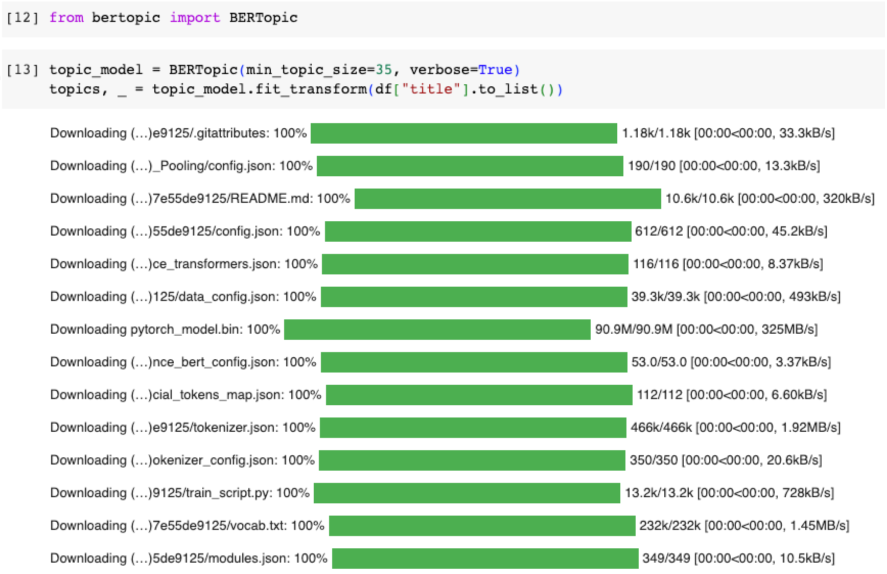


	doi	num_versions
0	10.1101/2020.07.09.20143164	26
1	10.1101/290825	25
2	10.1101/066423	19
3	10.1101/016840	19
4	10.1101/2020.05.26.104687	16
...	...	...

EDA: Topic analysis

BERTopic

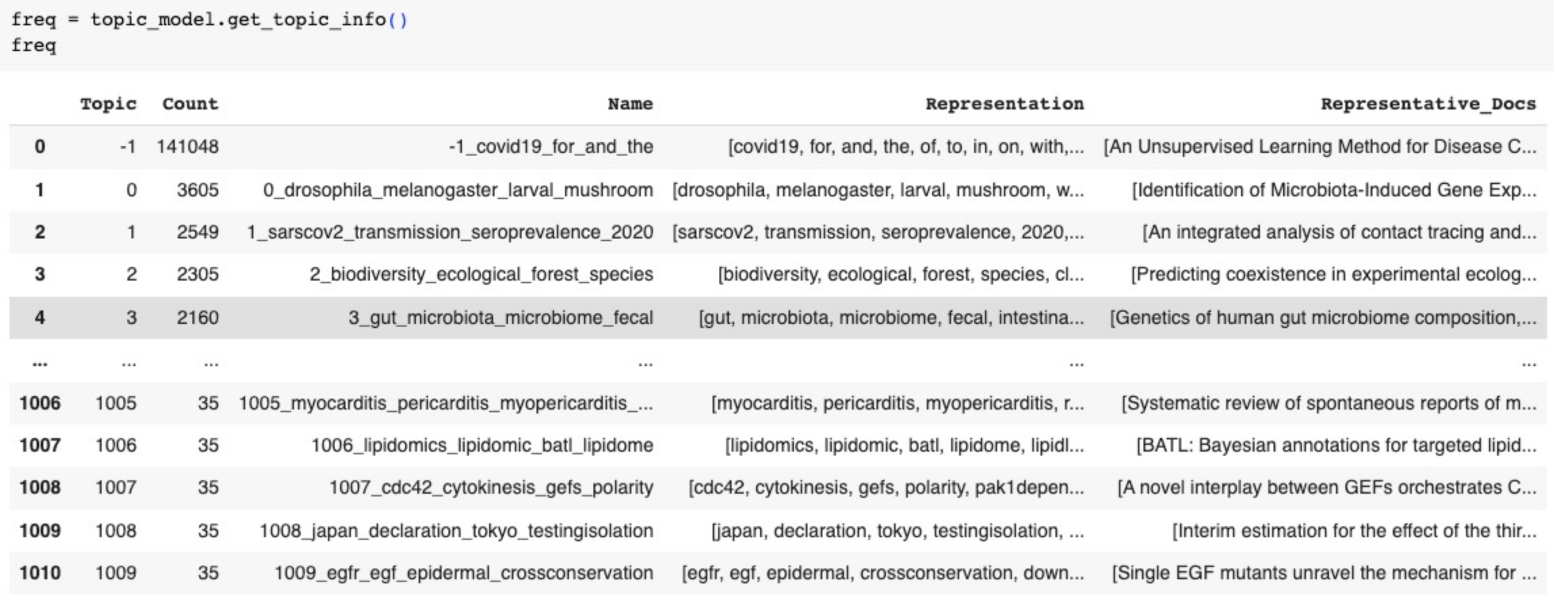
- Done on Google Colab in ~1 hour (20 mins init, 20 mins model fit, 20 mins analysis)
- Topics from titles
- No preprocessing



```
[12] from bertopic import BERTopic
```

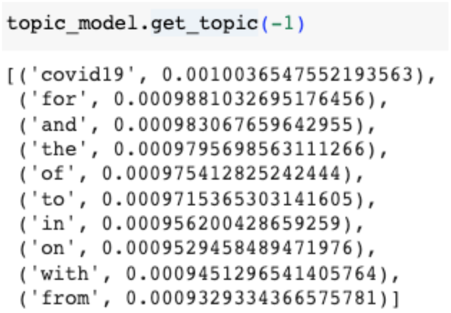
```
[13] topic_model = BERTopic(min_topic_size=35, verbose=True)
     topics, _ = topic_model.fit_transform(df["title"].to_list())
```

```
Downloading (…)e9125/.gitattributes: 100%  1.18k/1.18k [00:00<00:00, 33.3kB/s]
Downloading (…)_Pooling/config.json: 100%  190/190 [00:00<00:00, 13.3kB/s]
Downloading (…)7e55de9125/README.md: 100%  10.6k/10.6k [00:00<00:00, 320kB/s]
Downloading (…)55de9125/config.json: 100%  612/612 [00:00<00:00, 45.2kB/s]
Downloading (…)ce_transformers.json: 100%  116/116 [00:00<00:00, 8.37kB/s]
Downloading (…)125/data_config.json: 100%  39.3k/39.3k [00:00<00:00, 493kB/s]
Downloading pytorch_model.bin: 100%  90.9M/90.9M [00:00<00:00, 325MB/s]
Downloading (…)nce_bert_config.json: 100%  53.0/53.0 [00:00<00:00, 3.37kB/s]
Downloading (…)cial_tokens_map.json: 100%  112/112 [00:00<00:00, 6.60kB/s]
Downloading (…)e9125/tokenizer.json: 100%  466k/466k [00:00<00:00, 1.92MB/s]
Downloading (…)okenizer_config.json: 100%  350/350 [00:00<00:00, 20.6kB/s]
Downloading (…)9125/train_script.py: 100%  13.2k/13.2k [00:00<00:00, 728kB/s]
Downloading (…)7e55de9125/vocab.txt: 100%  232k/232k [00:00<00:00, 1.45MB/s]
Downloading (…)5de9125/modules.json: 100%  349/349 [00:00<00:00, 10.5kB/s]
```



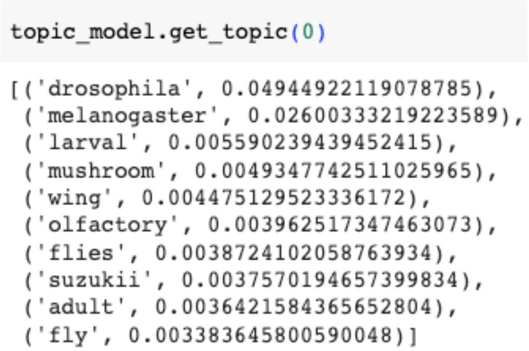
```
freq = topic_model.get_topic_info()
freq
```

	Topic	Count	Name	Representation	Representative_Docs
0	-1	141048	-1_covid19_for_and_the	[covid19, for, and, the, of, to, in, on, with,...	[An Unsupervised Learning Method for Disease C...
1	0	3605	0_drosophila_melanogaster_larval_mushroom	[drosophila, melanogaster, larval, mushroom, w...	[Identification of Microbiota-Induced Gene Exp...
2	1	2549	1_sarscov2_transmission_seroprevalence_2020	[sarscov2, transmission, seroprevalence, 2020,...	[An integrated analysis of contact tracing and...
3	2	2305	2_biodiversity_ecological_forest_species	[biodiversity, ecological, forest, species, cl...	[Predicting coexistence in experimental ecolog...
4	3	2160	3_gut_microbiota_microbiome_fecal	[gut, microbiota, microbiome, fecal, intestina...	[Genetics of human gut microbiome composition,...
...	...	...	...	...	...
1006	1005	35	1005_myocarditis_pericarditis_myopericarditis_...	[myocarditis, pericarditis, myopericarditis, r...	[Systematic review of spontaneous reports of m...
1007	1006	35	1006_lipidomics_lipidomic_batl_lipidome	[lipidomics, lipidomic, batl, lipidome, lipidl...	[BATL: Bayesian annotations for targeted lipid...
1008	1007	35	1007_cdc42_cytokinesis_gefs_polarity	[cdc42, cytokinesis, gefs, polarity, pak1depen...	[A novel interplay between GEFs orchestrates C...
1009	1008	35	1008_japan_declaration_tokyo_testingisolation	[japan, declaration, tokyo, testingisolation, ...	[Interim estimation for the effect of the thir...
1010	1009	35	1009_egfr_egf_epidermal_crossconservation	[egfr, egf, epidermal, crossconservation, down...	[Single EGF mutants unravel the mechanism for ...



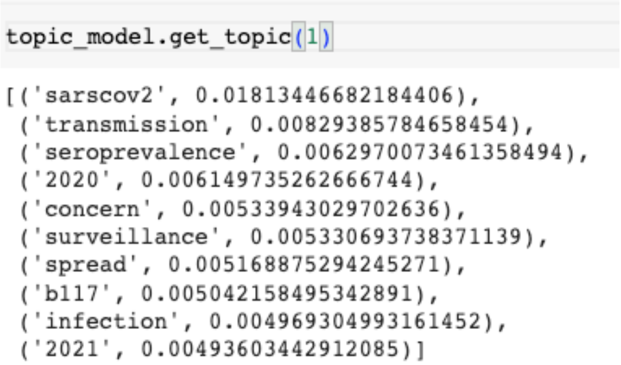
```
topic_model.get_topic(-1)
```

```
[('covid19', 0.0010036547552193563),
 ('for', 0.0009881032695176456),
 ('and', 0.000983067659642955),
 ('the', 0.0009795698563111266),
 ('of', 0.000975412825242444),
 ('to', 0.0009715365303141605),
 ('in', 0.000956200428659259),
 ('on', 0.0009529458489471976),
 ('with', 0.0009451296541405764),
 ('from', 0.0009329334366575781)]
```



```
topic_model.get_topic(0)
```

```
[('drosophila', 0.04944922119078785),
 ('melanogaster', 0.02600333219223589),
 ('larval', 0.005590239439452415),
 ('mushroom', 0.004934774251102596 5),
 ('wing', 0.004475129523336172),
 ('olfactory', 0.003962517347463073),
 ('flies', 0.003872410205876393 4),
 ('suzukii', 0.003757019465739983 4),
 ('adult', 0.003642158436565280 4),
 ('fly', 0.003383645800590048)]
```



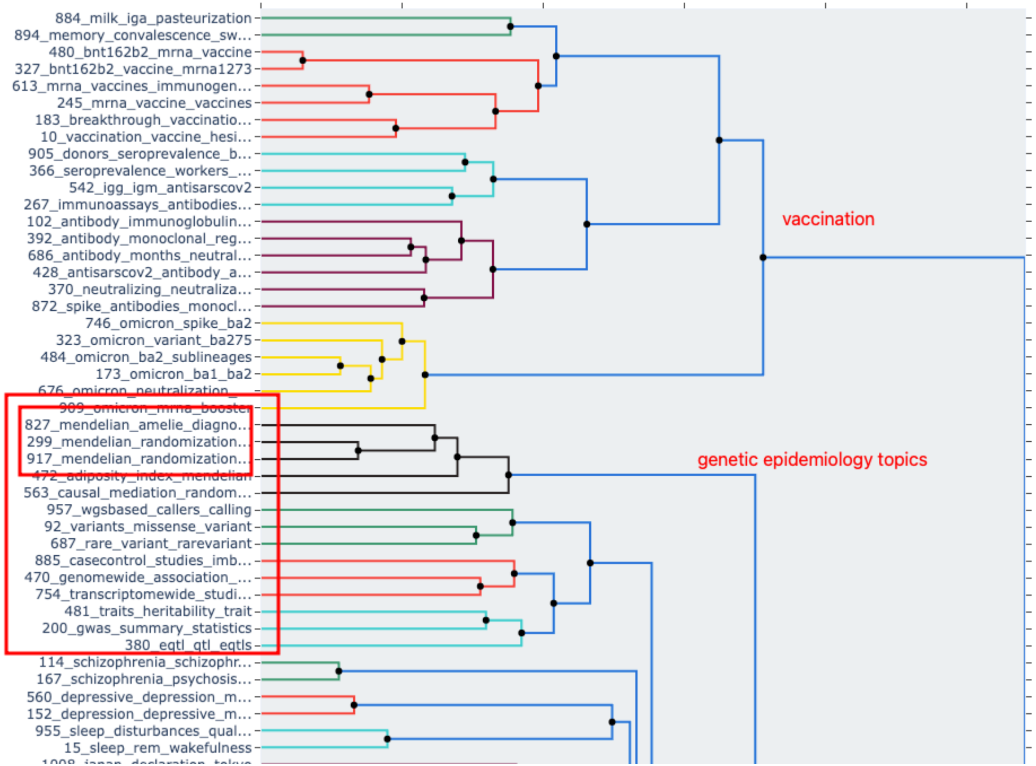
```
topic_model.get_topic(1)
```

```
[('sarscov2', 0.01813446682184406),
 ('transmission', 0.008293857846584 54),
 ('seroprevalence', 0.006297007346135849 4),
 ('2020', 0.006149735262666744),
 ('concern', 0.005339430297026 36),
 ('surveillance', 0.005330693738371139),
 ('spread', 0.005168875294245271),
 ('b117', 0.005042158495342891),
 ('infection', 0.004969304993161452),
 ('2021', 0.004936034429120 85)]
```

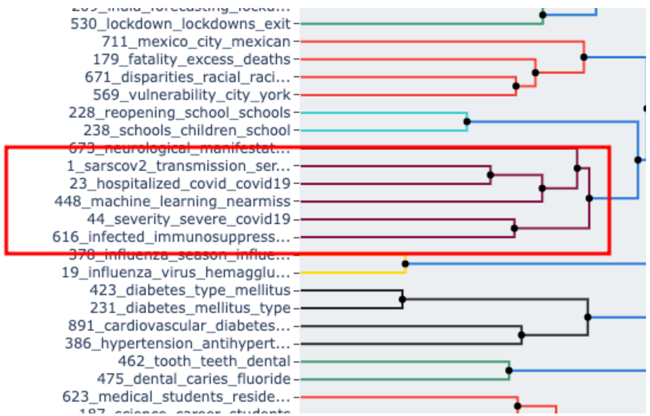
University of BRISTOL

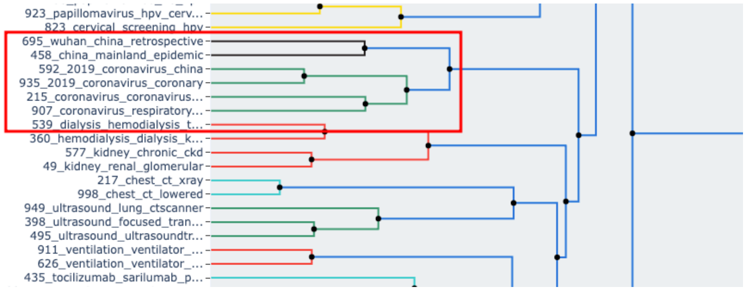
Elizabeth Blackwell Institute for Health Research

Hierarchical Clustering



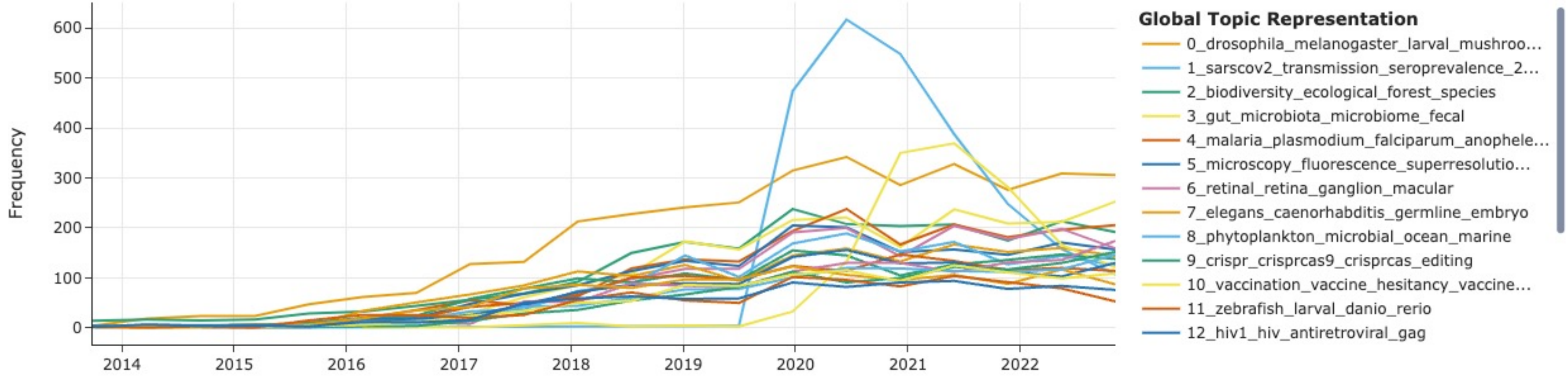
Covid severity
Coronovirus





bristol.ac.uk/blackwell @EBIBristol

Topics over Time



Next steps

- Write a blog about the project
- Put the text data into use, with collaboration projects
- Topics
  - Use the scientific text to train language models
  - Appropriate use of clustering methods to analyse research topics
  - What factors lead to successful publication of a biomedical preprint?
  - Assessment of risk-of-bias on preprints

- Questions, comments, suggestions welcome!

Acknowledgement

This work was supported by the Elizabeth Blackwell Institute for Health Research, University of Bristol, with funding from Research England’s quality-related (QR) award to the University of Bristol

This work was also carried out in the UK Medical Research Council Integrative Epidemiology Unit (MC_UU_00032/03).