Proceedings of the Biocomplexity Institute
Technical Report 2023-16

Detecting Pandemic Related R&D Trends using Dynamic Topic Modeling

Kathryn Linehan
https://orcid.org/0000-0001-9012-6261
kjl5t@virginia.edu

Guy Leonel SIWE
https://orcid.org/0000-0002-9275-6416
yhu2bk@virginia.edu

Joel Thurston
https://orcid.org/0000-0002-3923-9065
jt9sz@virginia.edu

Stephanie Shipp
https://orcid.org/0000-0002-2142-2136
sss5sc@virginia.edu

Audrey Kindlon
National Center for Science and Engineering Statistics

John Jankowski
National Center for Science and Engineering Statistics

Social and Decision Analytics Division
Biocomplexity Institute & Initiative
University of Virginia

January 31, 2023

Funding: This research was funded by the National Center for Science and Engineering Statistics (NCSES) contract 49100420C0015. The views expressed in this report are those of the authors and not the NCSES.

Citation: Linehan K, Siwe L, Thurston J, Shipp S, Kindlon A, Jankowski J. Detecting Pandemic Related R&D Trends using Dynamic Topic Modeling, Technical Report 2023-16. Proceeding of the Biocomplexity Institute, University of Virginia; 2023 January. Available from: https://doi.org/10.18130/zz1b-ww76.

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE

Abstract

This report explores the use of dynamic nonnegative matrix factorization (D-NMF) to identify topics and trends in federally funded R&D related to pandemics. We utilized scientific grant abstract data and compared D-NMF and static NMF topic model results on this data. We report topics discovered, topic prevalence, and provide an initial exploration into measures of D-NMF topic content change. While we found that D-NMF and static NMF identify many similar topics, we show that the D-NMF model can provide insightful supplementary results to those produced by the static NMF, and that the synthesis of results from D-NMF and static NMF can offer a more complete topic and trend analysis.

Detecting Pandemic Related R&D Trends using Dynamic Topic Modeling

Kathryn Linehan, Guy Leonel Siwe, Joel Thurston, Stephanie Shipp[1], Audrey Kindlon and John Jankowski[2]

Abstract

This report explores the use of dynamic nonnegative matrix factorization (D-NMF) to identify topics and trends in federally funded R&D related to pandemics. We utilized scientific grant abstract data from Federal RePORTER and compared D-NMF and static NMF topic model results on this data. We report topics discovered, topic prevalence, and provide an initial exploration into measures of D-NMF topic content change. While we found that D-NMF and static NMF identify many similar topics, we show that the D-NMF model can provide insightful supplementary results to those produced by the static NMF, and that the synthesis of results from D-NMF and static NMF can offer a more complete topic and trend analysis.

This research was funded by the National Center for Science and Engineering Statistics (NCSES) contract 49100420C0015. The views expressed in this report are those of the authors and not the NCSES.

[1] University of Virginia, Biocomplexity Institute, Social and Decision Analytics Division
[2] National Center for Science and Engineering Statistics

Detecting Pandemic Related R&D Trends using Dynamic Topic Modeling

Kathryn Linehan, Guy Leonel Siwe, Joel Thurston, Stephanie Shipp[3], Audrey Kindlon and John Jankowski[4]

1 Introduction

This report is an extension of previous work (Linehan et al., 2022) where we investigated the use of static nonnegative matrix factorization (NMF) to identify trends in federally funded research and development (R&D). The main purposes of this report are (1) to present our exploratory work on using the dynamic nonnegative matrix factorization (D-NMF) topic model to identify topics and trends in federally funded pandemic related R&D, and (2) to show examples where D-NMF provides useful supplementary information to that found through the use of static NMF. The dataset we use in this work is a subset of pandemic-related R&D grants funded in federal fiscal years (FYs) 2008-2020 from Federal RePORTER, a federal grant database (see Linehan et al. (2022) for more information). We identified this subset in previous research using term matching and latent semantic indexing (information on methods given in Linehan et al. (2022)).

While both D-NMF and static NMF create an overall list of topics for a set of documents, D-NMF can also show changes in topic content (the list of top words in a topic) over time. For example, while static NMF can discover a coronavirus topic in a corpus, D-NMF can show how coronavirus research has changed over time in that corpus. We explore this particular example in Section 5. The main takeaway from this work is that while D-NMF is more computationally intensive than that of static NMF, it can provide valuable information to supplement results of static NMF, and the synthesis of results from D-NMF and static NMF can offer a more complete topic and trend analysis. However, the degree of usefulness of the D-NMF results versus the computational trade-off must be determined by the scientist.

The remainder of this report is organized as follows, Section 2 presents brief examples of related work, Section 3 gives an overview of D-NMF, the dataset, and results of model parameter tuning, Section 4 includes D-NMF and static NMF results and discussion on federally funded pandemic related R&D topics and trends, Section 5 presents an example of D-NMF topic content change over time, and Section 6 concludes our work.

[3] University of Virginia, Biocomplexity Institute, Social and Decision Analytics Division

[4] National Center for Science and Engineering Statistics

2 Related Work

The two most commonly used dynamic topic models are dynamic latent Dirichlet allocation (Blei and Lafferty, 2006) and D-NMF (Greene and Cross, 2017). Dynamic topic models have been applied to detect changes in political agenda following internal and external stimuli (Greene and Cross, 2017), and study the time changes in topics from articles published in *Science* (Blei and Lafferty, 2006). They have also been used to capture changes in topic content covered in various text datasets such as personal emails, Neural Information Processing Systems (NeurIPS) papers, and presidential state-of-the-union addresses to describe how topic content changes relate to events (Wang and McCallum, 2006).

The D-NMF results discussed in this paper are related to the idea that major events can influence the direction of innovation. For example, multiple studies have shown a shift in the research focus from non-COVID 19 related research to COVID-19 related research following the COVID-19 pandemic (Raynaud et al., 2021; Riccaboni and Verginer, 2022). In addition, Bryan et al. (2020) showed this happened with a higher magnitude than research shifts following previous outbreaks such as Zika, Ebola, and H1N1. Other examples of a major event affecting innovation include penicillin advancements during World War II (World Intellectual Property Organization, 2022).

3 Methodology

3.1 D-NMF

We used D-NMF, developed by Greene and Cross (2017), with annual time periods to identify yearly topics and superordinate topics within a corpus of scientific grant abstracts. Each yearly topic can be linked to a superordinate topic, allowing us to track topic content changes over time. For example, suppose a superordinate cancer topic is linked to a yearly topic described by the words cancer, diagnostic, experiment, and a yearly topic described by the words cancer, vaccine, drugs in a later year. These changes in the topic content may represent a shift in cancer research from diagnostics to vaccines.

Input to D-NMF includes a set of documents that are processed, i.e., undergone standard natural language processing (NLP) steps such as tokenization, stop word removal, and lemmatization, and numerically represented as a TFIDF (term frequency-inverse document frequency) matrix. See our prior work for details (Linehan et al., 2022). D-NMF is based upon multiple NMFs. For topic modeling purposes, NMF approximately decomposes a TFIDF matrix, $\mathbf{M}$, into two matrices $\mathbf{W}$ and $\mathbf{H}$ for a user-chosen number of topics, k:

$$\underset{m\times n}{\mathbf{M}} \approx \underset{m\times k}{\mathbf{W}}\ \underset{k\times n}{\mathbf{H}}, \tag{1}$$

where m is the number of documents and n is the number of unique words in those documents. $\mathbf{W}$ is the document-topic matrix which gives the weight of each topic (in columns) for each document (in rows), and $\mathbf{H}$ is the topic-term matrix which gives the weight of each word (in columns) within each topic (in rows). A topic is characterized by its highest weighted words (usually top 5 or 10); these words should be related to a specific idea. To measure the performance of a topic model, C_V topic coherence can be used. C_V topic coherence measures how often the highest weighted words characterizing topics appear together across documents (Röder et al., 2015). The number of topics for an NMF model, k, can be chosen by selecting an initial range of potential values and then selecting the value for which C_V topic coherence is maximized.

D-NMF is a two stage model.

Stage One: Identification of Yearly Topics

Yearly topics refer to topics covered by documents from the same FY. For each FY, t,

1. select all documents from FY t,
2. build a TFIDF matrix, $\mathbf{M}_t$, from the selected documents, and
3. run an NMF model on $\mathbf{M}_t$ with k_t topics: $\mathbf{M}_t \approx \mathbf{W}_t\mathbf{H}_t$, where k_t is chosen to maximize C_V topic coherence as mentioned above.

This results in y NMF matrix factorizations that will be used in stage two, where y is the number of unique FYs for which there are documents.

Stage Two: Identification of Superordinate Topics

In stage two, the results from stage one are combined into a new matrix, $\mathbf{M}$ and an NMF is run on $\mathbf{M}$. This has the effect of clustering the yearly topics obtained from stage one into superordinate topics, where topics from different FYs that share similar ideas are clustered together.

1. Create a TFIDF-like matrix in which the yearly topics discovered in stage one are the "documents". This is achieved by concatenating all of the $\mathbf{H}_t$ (topic-term) matrices from stage one into a new matrix of size $\sum k_t$ by the number of unique words in all yearly topics. Call this matrix $\widehat{\mathbf{H}}$. This should be done so that one column of $\widehat{\mathbf{H}}$ corresponds to one word (where if the word does not correspond to a column of a particular $\mathbf{H}_t$, then the entries in $\widehat{\mathbf{H}}$ for this word and particular $\mathbf{H}_t$ will be zero).

2. For each row of $\widehat{\mathbf{H}}$, only keep the largest 10 entries. Set all other entries to zero. This step serves as feature selection and represents keeping only those words important to yearly topics.

3. Remove any columns of all zeros in $\widehat{\mathbf{H}}$ and call the resulting matrix $\mathbf{M}$. In effect, this removes any words that have not appeared in the top 10 of any yearly topics.

4. Run an NMF model on $\mathbf{M}$ with k topics: $\mathbf{M} \approx \mathbf{WH}$, where k is chosen to maximize C_V topic coherence as mentioned above.

In this case, $\mathbf{W}$ is the document-topic matrix which gives the weight of each superordinate topic (in columns) for each yearly topic (in rows), and $\mathbf{H}$ is the topic-term matrix which gives the weight of each word (in columns) within each superordinate topic (in rows). To identify superordinate topic content changes, each yearly topic is assigned to the superordinate topic for which it has the highest weight (i.e., the largest entry in the row of $\mathbf{W}$ corresponding to the yearly topic). A yearly topic from FY t assigned to a superordinate topic represents the content of that superordinate topic at time t.

3.2 Data

The dataset used in this work is a subset of pandemic-related R&D grants funded in FYs 2008-2020 from Federal RePORTER. This federal grant database includes grant titles, abstracts, funding agency, and FY, among other information. We identified this subset in previous research using term matching and latent semantic indexing (information on methods given in (Linehan et al., 2022)). Federal RePORTER includes information for more than 1.2 million grants. This "pandemics corpus" includes 7,571 grants.

Figure 1: Distribution of the pandemics corpus by FY and funding agency.

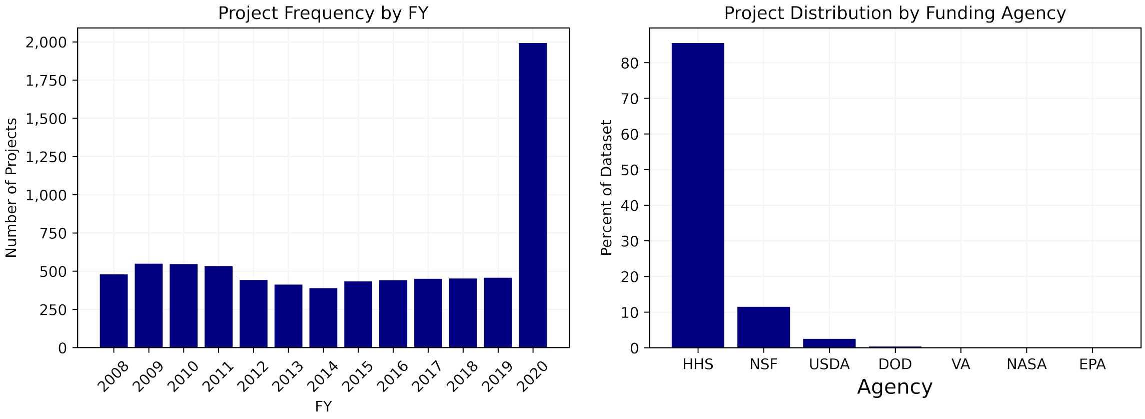


Figure (1) shows the distribution of the pandemics corpus by FY and funding agency. The number of grants related to pandemics is fairly steady (around 500) for FYs 2008-2019, with a large increase in FY 2020 likely driven by the COVID-19 pandemic. The US

Department of Health and Human Services (HHS) funds approxmately 85% of these grants and the National Science Foundation (NSF) funds approximately 10%.

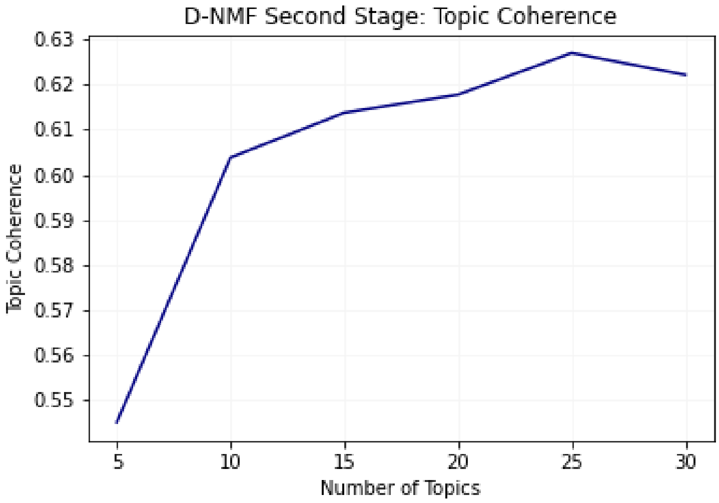
3.3 Model Parameter Tuning

To run D-NMF on the pandemics corpus, we identified k_t for each FY (i.e., the number of topics for each yearly topic model) in stage one, and k, the number of superordinate topics in stage two. Specifically, for each yearly topic model in stage one and the superordinate topic model in stage two, we utilized $\{5, 10, 15, ..., 35\}$ as potential values for the number of topics and chose the value that maximized the C_V topic coherence. Table (1) gives the results of the stage one tuning process, and Figure (2) presents results for the stage two tuning process, in which we see that 25 topics maximizes C_V topic coherence for the superordinate topic model.

Table 1: D-NMF stage one parameter tuning. The table presents the number of topics maximizing the C_V topic coherence and the resulting C_V topic coherence for each yearly topic model.

FY	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Number of Topics	10	30	25	20	25	25	20	20	10	10	15	30	30
Topic Coherence	0.626	0.608	0.648	0.640	0.626	0.650	0.657	0.667	0.642	0.630	0.651	0.636	0.646

Figure 2: D-NMF stage two parameter tuning. C_V topic coherence for various values of k, the number of topics, for the superordinate topic model.



4 Results and Discussion

In this section, the D-NMF results applied to the pandemics corpus are compared to the static NMF results applied to the same corpus. The D-NMF results include the yearly topics, superordinate topics, and a measure of prevalence for each topic. The static NMF results include a list of topics and a measure of prevalence for each topic. While static NMF on the pandemics corpus ran in 20 seconds, D-NMF ran in 418 seconds (almost 7 minutes).

4.1 Topics

In Table (2), D-NMF superordinate topics are given by their top 10 words ordered by weight, with the first word being the highest weighted. In Table (3), static NMF topics are given in the same format. The D-NMF superordinate topics include viruses such as influenza (topics 1 and 9), HIV (topics 3 and 5), and coronavirus (topic 23); and other topics such as vaccines (topic 4), public health risk information (topic 7), training programs (topic 8), and obesity (topic 25). The D-NMF superordinate topics and static NMF topics are similar.

Table 2: D-NMF superordinate topics for the pandemics corpus. The top 10 words characterizing each superordinate topic are given. Words are ordered from left to right by weight, with the first word listed having the highest weight in the topic.

Label	**Top Ten words**
Topic 1	influenza, vaccination, strain, effectiveness, ha, age, ve, household, estimate, universal
Topic 2	virus, human, animal, host, viral, evolution, genetic, 1918, disease, transmission
Topic 3	hiv, aids, prevention, trial, intervention, treatment, clinical, microbicide, woman, art
Topic 4	vaccine, protective, attenuate, adjuvant, efficacy, candidate, protection, strain, rsv, response
Topic 5	hiv_1, env, subtype, glycan, transmission, shiv, infection, recombinant, bnabs, sexual
Topic 6	cell, response, memory, lung, cd4, infection, immunity, immune, cd8, subset
Topic 7	risk, public, behavior, health, perception, information, policy, survey, response, people
Topic 8	training, program, trainee, student, faculty, infectious, university, health, disease, train
Topic 9	iav, lung, host, response, mast, infection, sp, evolution, mmp_9, evade
Topic 10	protection, immune, ecologic, pathogenicity, environmental, evolution, correlate, influence, animal, emergence
Topic 11	mtb, tb, ido, macaque, persister, granuloma, model, thiopeptide, co, aids
Topic 12	antibody, epitope, ha, neutralization, protective, neutralize, immunogen, escape, cross, conserve
Topic 13	drug, inhibitor, resistance, compound, antiviral, m2, resistant, treatment, protease, activity
Topic 14	siv, infection, co, malaria, transmission, model, infected, transmit_founder, mucosal, load
Topic 15	protein, viral, rna, host, interaction, replication, assembly, np, antiviral, cellular
Topic 16	dengue, serotype, wolbachia, virus, tetravalent, mosquito, flavivirus, vaccine, antibody, zika
Topic 17	zikv, zika, virus, mosquito, infection, wnv, cns, flavivirus, spread, microcephaly
Topic 18	ebola, virus, outbreak, infection, lassa, sudan, gp, ebov, compound, fusion
Topic 19	airway, gabaa, smooth_muscle, subunit, receptor, channel, asthma, agonist, relaxation, gaba
Topic 20	cancer, patient, care, kshv, treatment, breast, core, survivor, center, impact
Topic 21	child, family, parent, age, adult, rsv, school, language, year, population
Topic 22	bat, niv, rabie, human, transmission, bangladesh, outbreak, behavior, livestock, nipah_virus
Topic 23	sars_cov_2, infection, disease, covid_19, immune, severe, response, patient, lung, coronavirus
Topic 24	diagnostic, core, detection, technology, sample, poc, assay, rapid, device, lrs
Topic 25	obesity, insulin, metabolic, weight, gene, glucose, rygb, adipocyte, diabete, obese

Table 3: Static NMF topics for the pandemics corpus. The top 10 words characterizing each topic are given. Words are ordered from left to right by weight, with the first word listed having the highest weight in the topic.

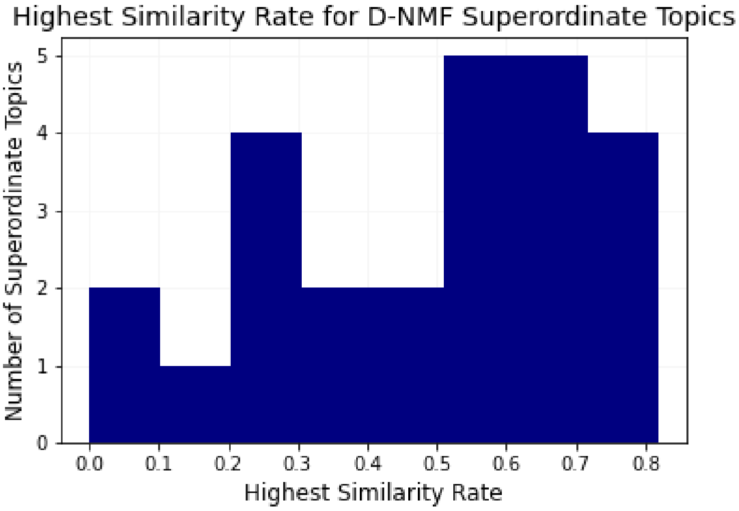
Label	Top Ten words
Topic 1	protein, rna, viral, structure, assembly, interaction, membrane, bind, np, target
Topic 2	covid_19, social, health, community, datum, risk, impact, survey, covid_19_pandemic, public
Topic 3	influenza, strain, virus, pandemic, vaccination, infection, ha, aim, death, antigenic
Topic 4	vaccine, adjuvant, candidate, efficacy, protection, protective, antigen, strain, attenuate, immunogenicity
Topic 5	hiv, aids, trial, prevention, intervention, infect, antiretroviral, clinical, infection, woman
Topic 6	cell, response, memory, cd4, infection, lung, immune, immunity, cd8, specific
Topic 7	protection, ecologic, immune, pathogenicity, environmental, correlate, evolution, cross, influence, emergence
Topic 8	hiv_1, env, subtype, transmission, infection, shiv, aids, glycan, infect, variant
Topic 9	training, program, student, trainee, faculty, career, university, mentor, doctoral, train
Topic 10	obesity, insulin, obese, metabolic, mouse, weight, increase, lipid, glucose, effect
Topic 11	drug, inhibitor, resistance, antiviral, compound, resistant, treatment, protease, target, therapeutic
Topic 12	iav, lung, sp, infection, mast, host, response, immunity, induce, evolution
Topic 13	antibody, epitope, ha, immunogen, neutralization, conserve, neutralize, human, protective, bind
Topic 14	diagnostic, detection, technology, sample, assay, rapid, core, device, poc, cost
Topic 15	sars_cov_2, covid_19, patient, infection, coronavirus, disease, severe, clinical, respiratory, syndrome
Topic 16	virus, human, infection, cause, replication, infect, attenuate, h5n1, avian_influenza, 1918
Topic 17	animal, bird, surveillance, close, contact, characterization, limit, rapid, serosurveillance, migratory_bird
Topic 18	dengue, serotype, virus, tetravalent, wolbachia, mosquito, denv, flavivirus, vaccine, infection
Topic 19	model, transmission, disease, intervention, spread, datum, epidemic, infectious, network, modeling
Topic 20	ebola, virus, outbreak, gp, filovirus, cycle, lassa, replication, entry, disease
Topic 21	host, evolution, viral, sequence, genetic, gene, genome, genomic, pathogen, diversity
Topic 22	tb, mtb, infection, co, disease, autophagy, tuberculosis_tb, treatment, progression, model
Topic 23	cancer, patient, kshv, care, treatment, breast, associate, center, aids, infection
Topic 24	zikv, zika, mosquito, wnv, cns, infection, flavivirus, microcephaly, pregnant, virus
Topic 25	rsv, child, age, ve, effectiveness, vaccine, adult, illness, year, respiratory

To quantify the amount of similarity between a D-NMF superordinate topic and a static NMF topic, we computed the Jaccard index, i.e., the fraction of top 10 words shared between them. All pairwise comparisons between D-NMF superordinate topics and static NMF topics are given in Table (4). Four pairs of topics achieve the highest similarity rate of 0.818:

- D-NMF superordinate topic 17 and static NMF topic 24 on Zika virus,
- D-NMF superordinate topic 24 and static NMF topic 14 on diagnostic testing,
- D-NMF superordinate topic 10 and static NMF topic 7 on ecologic immunity, and
- D-NMF superordinate topic 6 and static NMF topic 6 on memory T cells.

We note that 14 of the 25 superordinate D-NMF topics share at least half of their top 10 words with a static NMF topic. In addition, Figure (3) shows the distribution of D-NMF superordinate topics according to their highest fraction of top 10 words shared with a static NMF topic.

Figure 3: Distribution of D-NMF superordinate topics according to their highest similarity rate with a static NMF topic, i.e., the fraction of top 10 words shared with a static NMF topic.



4.2 Topic Prevalence

We followed the work of Greene and Cross (2017) to measure D-NMF topic prevalence by the number of abstracts associated to a topic. In the case of a D-NMF yearly topic, this is the number of abstracts with a non-zero weight for that topic. In the case of a D-NMF superordinate topic, this is the number of abstracts associated with at least one of the yearly topics assigned to the superordinate topic. (Each yearly topic is assigned to the superordinate topic for which it has the highest weight.) In order to compare topic prevalence between D-NMF and static NMF, we used the number of abstracts with a non-zero weight for a static NMF topic as its prevalence.

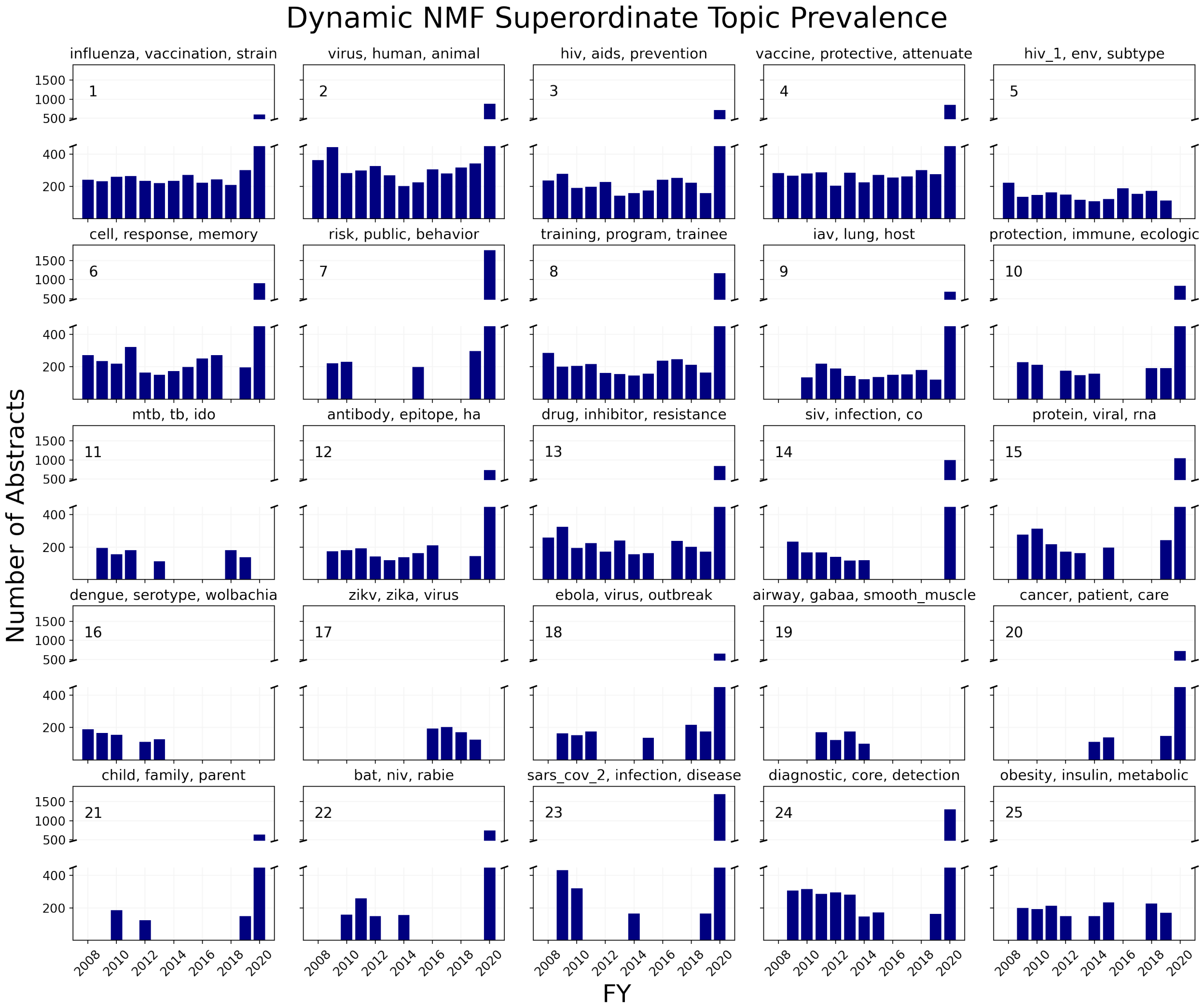
Although D-NMF and static NMF discover similar topics, they provide different trends in topic prevalence. Figures (4) and (5) show the prevalence of each D-NMF superordinate topic and each static NMF topic respectively. The static NMF results show abstracts associated with each topic every FY; this is due to how the number of abstracts associated with each static NMF topic is calculated. Even an abstract with a very small weight for a static NMF topic will be associated to that topic. There is also a large increase in the number of grants in FY 2020 for all static NMF topics, likely due to the surge of COVID-19 related research. We also note that all static NMF topics show an increase in prevalence around FYs 2009-2010, likely due to the American Recovery and Reinvestment Act of 2009 (ARRA).

In comparison, the D-NMF results provide a different perspective on topic prevalence since yearly topics are only assigned to the superordinate topic for which they have

Table 4: Similarity rate between each D-NMF superordinate topic with each static NMF topic. The similarity rate between two topics is the Jaccard index, i.e., the fraction of top 10 words shared between those topics.

		Static NMF																								
Topic number		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
D-NMF	**1**	0	0	0.25	0.053	0	0	0	0	0	0	0	0	0.053	0	0	0	0	0	0	0	0	0	0	0	0.176
	2	0.053	0	0.053	0	0	0	0.053	0.053	0	0	0	0.111	0.053	0	0.053	0.176	0.053	0.053	0.111	0.111	0.25	0.053	0	0.053	0
	3	0	0	0	0	0.538	0	0	0.053	0	0	0.053	0	0	0	0.053	0	0	0	0.053	0	0	0.053	0.111	0	0
	4	0	0	0.053	0.667	0	0.053	0.053	0	0	0	0	0.053	0.053	0	0	0.053	0	0.053	0	0	0	0	0	0	0.111
	5	0	0	0.053	0	0.053	0.053	0	0.538	0	0	0	0.053	0	0	0.053	0.053	0	0.053	0.053	0	0	0.053	0.053	0.053	0
	6	0	0	0.053	0	0.053	0.818	0.053	0.053	0	0	0	0.25	0	0	0.053	0.053	0	0.053	0	0	0	0.053	0.053	0.053	0
	7	0	0.25	0	0	0	0.053	0	0	0	0	0	0.053	0	0	0	0	0	0	0	0	0	0	0	0	0
	8	0	0.053	0	0	0	0	0	0	0.538	0	0	0	0	0	0.053	0	0	0	0.111	0.053	0	0.053	0	0	0
	9	0	0	0.053	0	0.053	0.176	0.053	0.053	0	0	0	0.667	0	0	0.053	0.053	0	0.053	0	0	0.111	0.053	0.053	0.053	0
	10	0	0	0	0.053	0	0.053	0.818	0	0	0	0	0.053	0	0	0	0	0.053	0	0	0	0.053	0	0	0	0
	11	0	0	0	0	0.053	0	0	0.053	0	0	0	0	0	0	0	0	0	0	0.053	0	0	0.25	0.053	0	0
	12	0	0	0.053	0.053	0	0	0.053	0	0	0	0	0	0.667	0	0	0	0	0	0	0	0	0	0	0	0
	13	0	0	0	0	0	0	0	0	0	0	0.667	0	0	0	0	0	0	0	0	0	0	0.053	0.053	0	0
	14	0	0	0.053	0	0.053	0.053	0	0.111	0	0	0	0.053	0	0	0.053	0.053	0	0.053	0.111	0	0	0.176	0.053	0.053	0
	15	0.429	0	0	0	0	0	0	0	0	0	0.053	0.053	0	0	0	0.053	0	0	0	0.053	0.111	0	0	0	0
	16	0	0	0.053	0.053	0	0	0	0	0	0	0	0	0.053	0	0	0.053	0	0.667	0	0.053	0	0	0	0.25	0.053
	17	0	0	0.111	0	0.053	0.053	0	0.053	0	0	0	0.053	0	0	0.053	0.111	0	0.25	0.053	0.053	0	0.053	0.053	0.818	0
	18	0	0	0.111	0	0.053	0.053	0	0.053	0	0	0.053	0.053	0	0	0.053	0.111	0	0.111	0	0.333	0	0.053	0.053	0.111	0
	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	20	0	0.053	0	0	0	0	0	0	0	0	0.053	0	0	0.053	0.053	0	0	0	0	0	0	0.053	0.538	0	0
	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.333
	22	0	0	0	0	0	0	0	0.053	0	0	0	0	0.053	0	0	0.053	0	0	0.053	0.053	0	0	0	0	0
	23	0	0.053	0.053	0	0.053	0.25	0.053	0.053	0	0	0	0.176	0	0	0.538	0.053	0	0.053	0.053	0.053	0	0.111	0.111	0.053	0
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0.818	0	0	0.053	0	0	0	0	0	0	0	0
	25	0	0	0	0	0	0	0	0	0	0.429	0	0	0	0	0	0	0	0	0	0	0.053	0	0	0	0

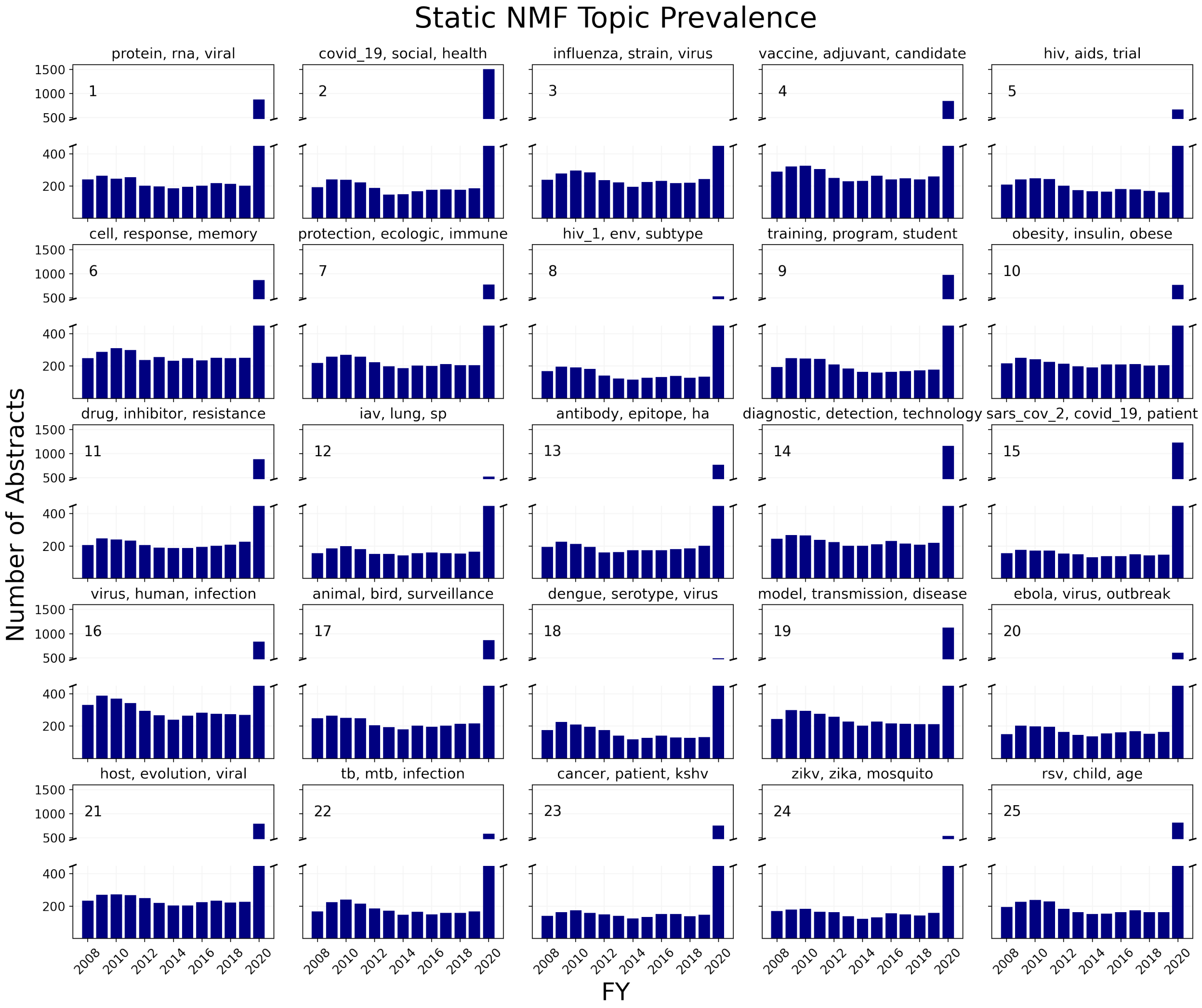
Figure 4: Pandemics corpus D-NMF superordinate topic prevalence over time as given by the number of abstracts associated with each topic in each FY. Each subplot corresponds to a superordinate topic; the title is the top three topic terms, and the number in the upper left hand corner matches the topic label number from Table (2).



the highest weight. Hence, most D-NMF superordinate topics do not have associated abstracts in every FY. However, this can actually provide useful information that the static NMF topic prevalence does not. For example, D-NMF superordinate topic 17 and static NMF topic 24 are both about Zika virus and are very similar (they share 81.8% of their top ten terms). The static NMF topic 24 prevalence is relatively steady for FYs 2008-2019. However, the D-NMF superordinate topic 17 prevalence shows that there are only abstracts associated to this topic in FYs 2016-2019, which follows the 2015-2016 Zika outbreak.

For a topic that persists over FYs 2008-2020, the D-NMF superordinate topic prevalence can show a similar trend to that of the static NMF topic prevalence. For example, D-NMF superordinate topic 8 and static NMF topic 9 on training programs share 53.8% of their top ten terms so are relatively similar. While the static NMF topic 9 prevalence has

Figure 5: Pandemics corpus static NMF topic prevalence over time as given by the number of abstracts associated with each topic in each FY. Each subplot corresponds to a topic; the title is the top three topic terms, and the number in the upper left hand corner matches the topic label number from Table (3).



less variance over FYs 2008-2019 than the D-NMF superordinate topic 8 prevalence, each prevalence measure is about 200 in every FY between 2008-2019.

In addition, the D-NMF superordinate topics with associated abstracts in FY 2020 show a large increase in prevalence over previous years, likely due to COVID-19 research. The static NMF topic prevalence results also showed this trend. However, the increase in topic prevalence due to ARRA in FYs 2009-2010 as seen in the static NMF topic prevalence results is not visible in the D-NMF superordinate topic prevalence results. Hence, using both D-NMF and static NMF results provides a more complete analysis of pandemics R&D topics and trends.

5 Coronavirus Example

In this section, we show an example of D-NMF superordinate topic content change over time. This is information that static NMF cannot provide. We focus on D-NMF superordinate topic 23 about coronavirus and its assigned yearly topics as given in Table (5), which provide information about how research related to this superordinate topic has changed over time. In FY 2009, there are yearly topics about five different viruses; all are respiratory viruses except cholera, however only one of these (Severe Acute Respiratory Syndrome - SARS) is a coronavirus. In FY 2010, there are yearly topics about respiratory infections and cholera; in FY 2014 there is one yearly topic about pneumonia and Middle East Respiratory Syndrome (MERS), another coronavirus; and in FY 2019 there is another topic on respiratory infection and pneumonia. However, in FY 2020 there are yearly topics on COVID-19 infection; COVID-19 testing, intervention and social disparity; and alcohol use related to the COVID-19 pandemic. While the yearly topics in FYs 2009, 2010, 2014, and 2019 covered a variety of viruses, they were all respiratory viruses except for two cholera topics (one in FY 2009 and one in FY 2010). In FY 2020, we observe the shift in focus to COVID-19, while simultaneously capturing a number of distinct elements of interest such as community testing versus risk factors (e.g., alcohol use).

Table 5: D-NMF yearly topics assigned to D-NMF superordinate topic 23. The top 10 words characterizing each topic are given. Words are ordered from left to right by weight, with the first word listed having the highest weight in the topic.

Superordinate Topic 23		sars_cov_2, infection, disease, covid_19, immune, severe, response, patient, lung, coronavirus
Yearly Topics	**2009**	v_cholarea, cholera, colonization, virulence, bacterial, lps, lipid, strain, disease, host
		sars_cov, coronaviruse, pathogenesis, vlp, sars, attenuate, trn, coronavirus, respiratory, orf
		cmv, hcmv, challenge, immunization, titer, infection, vaccinee, immune, protective, rhcmv
		h5n1, h5n1_influenza, infection, poultry, worker, human, asia, transmission, influenza, pandemic
		prrs, swine, disease, reproductive, industry, pig, pork, cause, economic, respiratory
	2010	infection, lung, immune, response mouse, inflammatory, induce, innate, cytokine, host
		v_cholerae, cholera, colonization, lps, virulence, lipid, disease, bacterial, environment, vibrio
	2014	lung, infection, pneunomia, disease, model, mers_cov, animal, mouse, human, pvl
	2019	lung, ds, infection, ifn, socs1, bacterial, pneumonia, coinfection, susceptibility, alveolar
	2020	patient, covid_19, disease, clinical, severe, risk, severity, infection, outcome, hospitalize
		community, testing, rural, covid_19, health, intervention, population, disparity, undeserved, county
		sars_cov_2, infection, disease, response, immune, coronavirus, respiratory, severe, covid_19, syndrome
		alcohol, use, consumption, aud, stress, relate, misuse, covid_19, risk, drinking

6 Conclusion

In this work we explored the use of D-NMF on scientific grant abstract data. Specifically, we used federally funded pandemic-related R&D grant abstracts and ran D-NMF and static NMF on this corpus to discover topics and trends. We showed that while many of the D-NMF superordinate topics are similar to the static NMF topics, we do gain additional insights from the D-NMF superordinate topic prevalence and by leveraging the assigning of D-NMF yearly topics to D-NMF superordinate topics. We showed the latter through an example in which we presented the evolution of a D-NMF superordinate topic about coronavirus over time. However, D-NMF does take much longer to run than static NMF; on our small pandemics corpus it took about 21 times longer. This is a limiting factor for D-NMF and may prevent its use on larger datasets. Nevertheless, for a dataset on which D-NMF is computationally feasible, D-NMF can provide insightful supplemental information to that discovered by static NMF, and the synthesis of results from D-NMF and static NMF can offer a more complete topic and trend analysis.

7 Acknowledgements

We would like to thank Eric Oh, former Research Assistant Professor at the Social and Decision Analytics Division, University of Virginia Biocomplexity Institute, for initial contributions to this work in running D-NMF on the pandemics corpus.

References

Blei, D. M. and Lafferty, J. D. (2006). Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. Association for Computing Machinery.

Bryan, K., Lemus, J., and Marshall, G. (2020). Innovation During a Crisis: Evidence from Covid-19. Available at SSRN: `https://ssrn.com/abstract=3587973` or `http://dx.doi.org/10.2139/ssrn.3587973`.

Greene, D. and Cross, J. P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*, 25(1):77–94.

Linehan, K., Oh, E., Thurston, J., Siwe, G. L., Garrett, M., Keller, S., Shipp, S., Kindlon, A., and Jankowski, J. (2022). Technical Report - Detecting Federally Funded Research and Development Trends Using Machine Learning and Information Retrieval Methods. Technical Report BI-2022-1531, Proceedings of the Biocomplexity Institute. University of Virginia.

Raynaud, M., Goutaudier, V., Louis, K., Al-Awadhi, S., Dubourg, Q., Truchot, A., Brousse, R., Saleh, N., Giarraputo, A., Debiais, C., Demir, Z., Certain, A., Tacafred, F., Cortes-Garcia, E., Yanes, S., Dagobert, J., Naser, S., Robin, B., Bailly, E., Jouven, X., Reese, P. P., and Loupy, A. (2021). Impact of the COVID-19 pandemic on publication dynamics and non-COVID-19 research production. *BMC Medical Research Methodology*, 21:Article 255.

Riccaboni, M. and Verginer, L. (2022). The impact of the COVID-19 pandemic on scientific research in the life sciences. *PLOS ONE*, 17(2):e0263001.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Wang, X. and McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 424–433, New York, NY, USA. Association for Computing Machinery.

World Intellectual Property Organization (2022). *World Intellectual Property Report 2022: The direction of innovation*. Geneva: WIPO.

Appendix A Measures for D-NMF Superordinate Topic Content Change

One main feature of D-NMF is that it can track changes in superordinate topic content through assigned yearly topics. For example, Table (5) reports the changes in content over time of superordinate topic 23 about coronavirus, described by its assigned yearly topics. Greene and Cross (2017) quantified the content change of a superordinate topic by computing the mean "Jaccard agreement" between each pair of consecutive yearly topics assigned to the superordinate topic. The Jaccard agreement between two yearly topics is the Jaccard index between the two sets of the top w words for each yearly topic, i.e.,

$$J(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|},$$

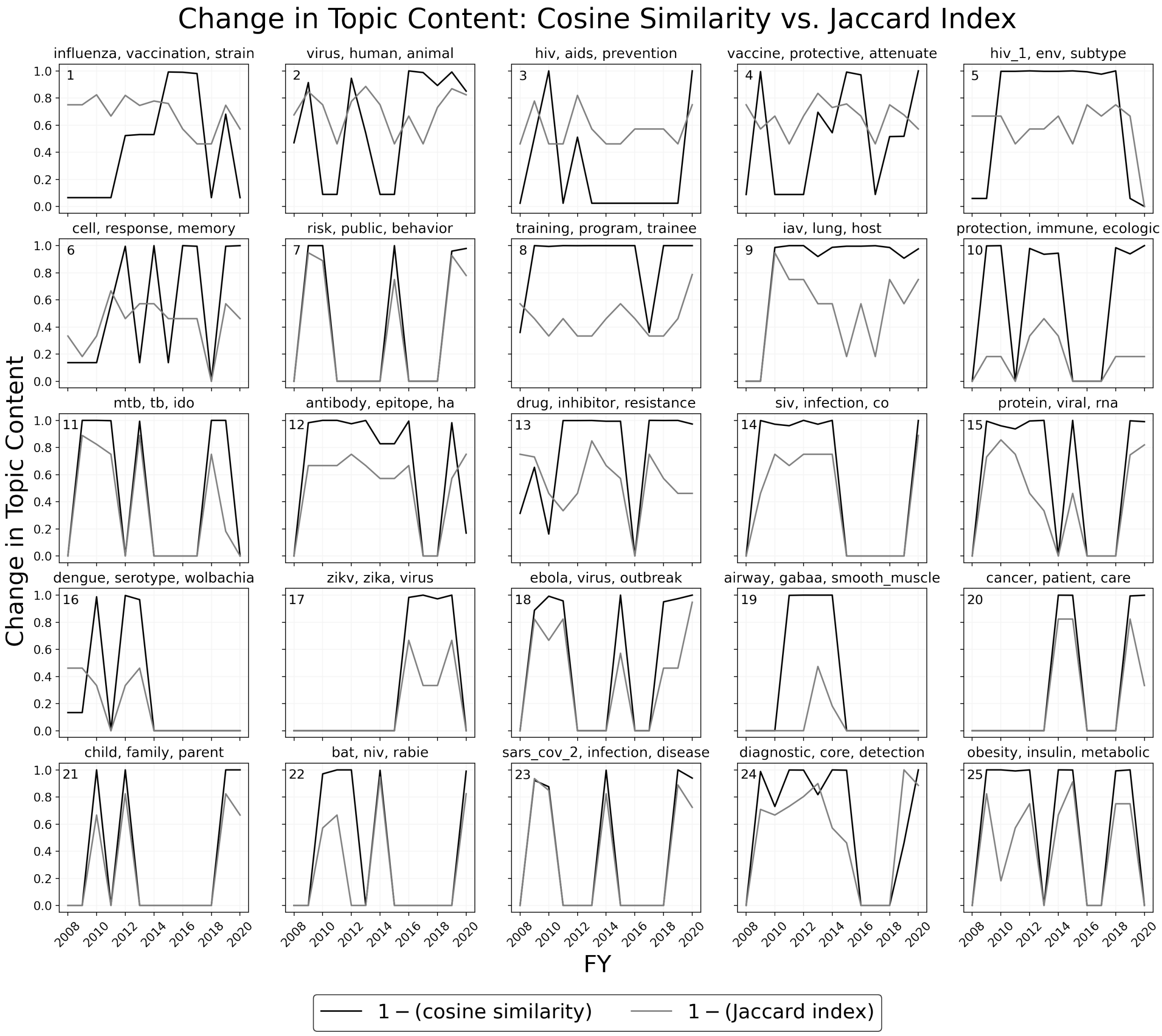
where T_1 and T_2 are the sets of top w words for each yearly topic. (In other words, the Jaccard index computes the fraction of top words shared by two topics.) In this appendix we present two additional metrics for quantifying superordinate topic content change. This exploration was motivated by the fact that most of our pandemics corpus D-NMF superordinate topics do not have abstracts associated with them in every FY (see Figure (4)). For example, superordinate topic 23 (coronavirus) has abstracts associated with it in FYs 2009, 2010, 2014, 2019, and 2020. Hence, we wanted to explore superordinate topic content change measures that we could track over time, rather than those that produce a single score.

Thus, to measure the content change of a superordinate topic, we calculated 1) the mean Jaccard index between the superordinate topic and each of its assigned yearly topics (using the top ten words from each respective topic), and 2) the mean cosine similarity between the vector representations of the superordinate topic and each of its assigned yearly topics. The vector representation of the superordinate topic is a result of stage two of D-NMF without the feature selection process (i.e., all terms are kept for each yearly topic), while the vector representation of each yearly topic is a result of stage one of D-NMF. These two measures differ in the fact that the Jaccard index captures variations in the top ten words characterizing a topic, whereas the cosine similarity considers variations in the weights of all topic words.

We subtracted each measure from one so that a score of zero for each measure signals no difference in content between the superordinate topic and its assigned yearly topic(s), and a score of one signals a large difference in content between the superordinate topic and its assigned yearly topic(s). In addition, the score for each measure will be zero for any FYs in which there are no yearly topics assigned to the superordinate topic. In Figure (6) we present both measures over time for every superordinate topic of the D-NMF that we ran on the pandemics corpus.

For some superordinate topics such as topic 23 (coronavirus) and topic 20 (cancer), the two measures are similar over time; however this does not necessarily need to be the case. For example, for other superordinate topics such as topic 1 (influenza) and topic 3 (HIV/AIDS), the two measures are not similar, reflecting the type of content change observed: a change in the top ten topic words (measured by the Jaccard index) or a change in the weights of all topic words (measured by the cosine similarity). This appendix presented initial steps into researching alternative measures for quantifying D-NMF superordinate topic content change. Expanding upon this could be future work.

Figure 6: Pandemics corpus D-NMF superordinate topic content change over time. Each subplot corresponds to a superordinate topic; the title is the top three topic terms, and the number in the upper left hand corner matches the topic label number from Table (2).



About the University of Virginia’s Social and Decision Analytics Division

The **Social and Decision Analytics Division (SDAD)** is a leading Division in the Biocomplexity Institute at the University of Virginia. The Biocomplexity Institute is at the forefront of a scientific evolution, applying a deeply contextual approach to answering some of the most pressing challenges to human health and well-being within our changing environment. SDAD was created in the fall of 2013 to extend the Biocomplexity Institute’s capabilities in social informatics, policy analytics, and program evaluation. The researchers at SDAD form a multidisciplinary team, with expertise in statistics, policy and program evaluation, economics, political science, psychology, computational social science, and data governance and information architecture. SDAD’s mission is to embrace today’s data revolution, developing evidence-based research and quantitative methods to inform policy decision-making and evaluation.