

SyS CGI: Novel Methods for SyS Practice

2:00 PM EST | December 7th, 2022 | 2022 SyS Symposium

Moderator:

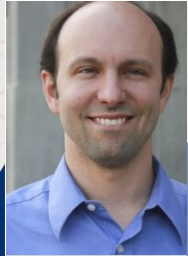
- Howard Burkom, PhD – Johns Hopkins Applied Physics Laboratory (JHUAPL)

Presenters:

- Daniel B. Neill, PhD – New York University
- Michael Sheppard, MS – CDC



Council of State and Territorial Epidemiologists



Daniel B. Neill, PhD
New York University



Pre-syndromic surveillance for improved detection of emerging public health threats

Daniel B. Neill, Ph.D.

**Machine Learning for Good Laboratory
New York University**

daniel.neill@nyu.edu

Joint work with Mallory Nobles (CMU), Ramona Lall (NYC DOHMH), and Rob Mathes (NYC DOHMH).



NYU

Center for Urban
Science + Progress

Machine Learning
for Good Laboratory

EPIDEMIOLOGY

Presyndromic surveillance for improved detection of emerging public health threats

Mallory Nobles^{1†}, Ramona Lall², Robert W. Mathes², Daniel B. Neill^{3*}

Existing public health surveillance systems that rely on predefined symptom categories, or syndromes, are effective at monitoring known illnesses, but there is a critical need for innovation in “presyndromic” surveillance that detects biothreats with rare or previously unseen symptomology. We introduce a data-driven, automated machine learning approach for presyndromic surveillance that learns newly emerging syndromes from free-text emergency department chief complaints, identifies localized case clusters among subpopulations, and incorporates practitioner feedback to automatically distinguish between relevant and irrelevant clusters, thus providing personalized, actionable decision support. Blinded evaluations by New York City’s Department of Health and Mental Hygiene demonstrate that our approach identifies more events of public health interest and achieves a lower false-positive rate compared to a state-of-the-art baseline.

INTRODUCTION

To offer a rapid, targeted, and effective response to emerging biothreats, public health officials must be able to detect a huge variety of emerging events. Recent, high-profile events highlight the diversity of situations that can affect public health: In February 2020, 50+ residents of a nursing home in Kirkland, Washington were part of one of the first coronavirus disease 2019 (COVID-19) outbreaks in the

can result in substantial detection delays. Moreover, syndromic surveillance can dilute the signal of a rare outbreak or novel biothreat, either by grouping rare cases with more common illnesses, or by splitting cases among many syndromes. In either case, the syndromic surveillance system may require a large increase in cases to recognize an anomalous cluster corresponding to a rare or novel event, making it difficult for public health to achieve timely detection and response.

Copyright © 2022
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim to
original U.S. Government
Works. Distributed
under a Creative
Commons Attribution
NonCommercial
License 4.0 (CC BY-NC).

This presentation is based on
our recent publication:

M. Nobles, R. Lall, R.W.
Mathes, and D.B. Neill*,
Science Advances 8(44):
eabm4920, 2022.

DOI: 10.1126/sciadv.abm4920
(open access)

Acknowledgements:

This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330 and the DHS Hidden Signals Challenge.

Thanks to the BCD Syndromic Surveillance Unit at NYC DOHMH for providing retrospective data for this study and participating in blinded evaluations, and to Boyuan Chen and Andy Wei for their work in developing the open-source Python version of MUSES.

Pre-syndromic surveillance

<u>Date/time</u>	<u>Hosp.</u>	<u>Age</u>	<u>Complaint</u>
Jan 1 08:00	A	19-24	runny nose
Jan 1 08:15	B	10-14	fever, chills
Jan 1 08:16	A	0-1	broken arm
Jan 2 08:20	C	65+	vomited 3x
Jan 2 08:22	A	45-64	high temp

Key challenge: A syndrome cannot be created to identify every possible cluster of potential public health significance.

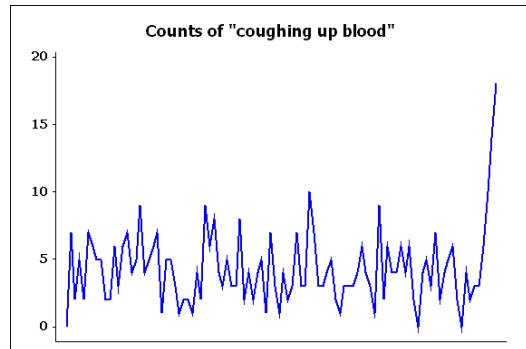
Thus a method is needed to identify relevant clusters of disease cases that do not correspond to existing syndromes.

Use case proposed by NC DOH and NYC DOHMH, solution requirements developed through a public health consultancy at the International Society for Disease Surveillance.

Where do existing methods fail?

The typical syndromic surveillance approach can effectively detect emerging outbreaks with commonly seen, general patterns of symptoms (e.g. ILI).

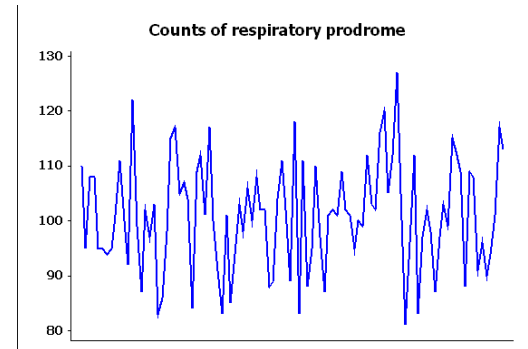
If we were monitoring these particular symptoms, it would only take a few such cases to realize that an outbreak is occurring!



What happens when something new and scary comes along?

- **More specific symptoms**
("coughing up blood")
- **Previously unseen symptoms**
("nose falls off")

Mapping specific chief complaints to a broader symptom category can dilute the outbreak signal, delaying or preventing detection.

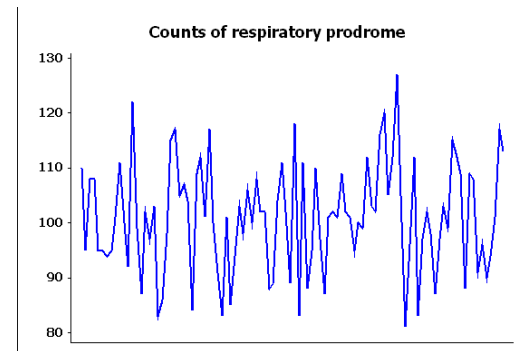
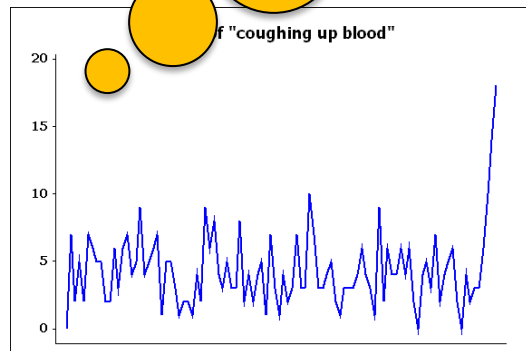


Where do existing methods fail?

The typical surveillance system is designed to detect something along? symptoms (d") on (off") effectively outbreak seen sy

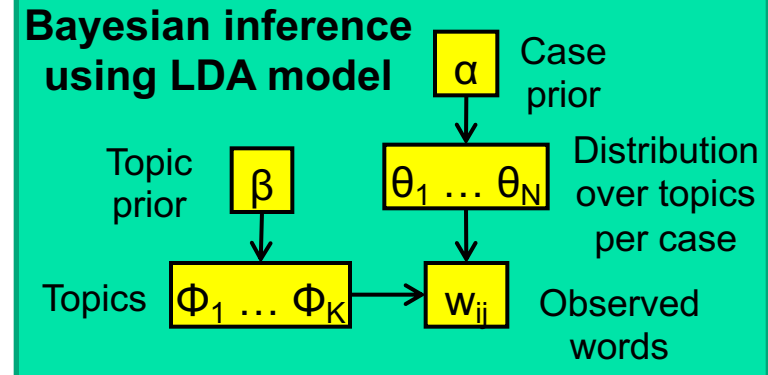
Our solution is to combine text-based (topic modeling) and event detection (multidimensional scan) approaches, to detect **emerging patterns of keywords.**

If we w particular sy take a few such that an outbreak is occurring! brief complaints symptom category make the outbreak signal, saying or preventing detection.



The semantic scan statistic

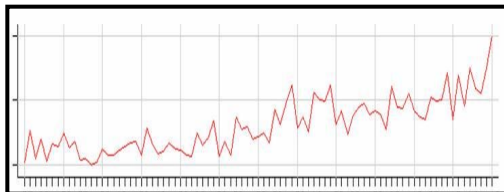
<u>Date/time</u>	<u>Hosp.</u>	<u>Age</u>	<u>Complaint</u>
Jan 1 08:00	A	19-24	runny nose
Jan 1 08:15	B	10-14	fever, chills
Jan 1 08:16	A	0-1	broken arm
Jan 2 08:20	C	65+	vomited 3x
Jan 2 08:22	A	45-64	high temp



Classify cases to topics



ϕ_1 : vomiting, nausea, diarrhea, ...
 ϕ_2 : dizzy, lightheaded, weak, ...
 ϕ_3 : cough, throat, sore, ...



Time series of hourly counts for each combination of hospital and age group, for each topic ϕ_j .

Now we can do a multidimensional scan, using the learned topics instead of pre-specified syndromes!

NYC DOHMH dataset

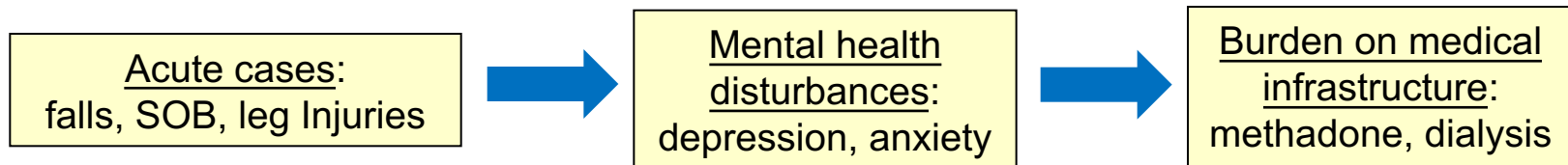
- New York City's Department of Health and Mental Hygiene, Bureau of Communicable Disease, provided us with 6 years of data (2010-2016) consisting of ~28M chief complaint cases from 53 hospitals in NYC.
- For each case, we have data on the patient's chief complaint (free text), date and time of arrival, age group, gender, and discharge ICD code.
- Substantial pre-processing of the chief complaint field was necessary because of the size and messiness of the data (typos, abbreviations, etc.).

VOIMITING	VOMITINIG	VOMITINGN
VOIMITTING	VOMITINNG	VOMITINGQ
VOIMTING	VOMITIONG	VOMITINGS
VOMIITING	VOMITITING	VOMITINGT
VOMIITNG	VOMITITNG	VOMITINGX
VOMINITING	VOMITN	VOMITINGX1
VOMINTING	VOMITNG	VOMITINGX2
VOMIOTING	VOMITNIG	VOMITINGX3
VOMITE	VOMITNING	VOMITINGX4
VOMITED	VOMITO	VOMMITTING
VOMITG	VOMITOS	VOMNITING
VOMITHING	VOMITS	VOMOITING
VOMITI	VOMITT	VOMTIING
VOMITIG	VOMITTE	VOMTIN
VOMITIGN	VOMITTI	VOMTITING
VOMITIING	VOMITTING	VONMITING
VOMITIN	VOMITTTING	VOOMITING
VOMITING3	VOMITUS	VOPMITING
VOMITINGA	VOMMIT	VVOMITING
VOMITINGG	VOMMITING	VOMITINGM

Variations of the words "vomit" and "vomiting" that appear > 15 times in data

Events identified by semantic scan

The progression of detected clusters after Hurricane Sandy impacted NYC highlights the variety of strains placed on hospital emergency departments following a natural disaster:



Many other events of public health interest were identified:

Accidents
Motor vehicle
Ferry
School bus
Elevator

Contagious Diseases
Meningitis
Scabies
Ringworm
Hepatitis

Other
Drug overdoses
Smoke inhalation
Carbon monoxide poisoning
Crime related, e.g., pepper spray attacks

Example of a detected cluster

Arrival Date	Arrival Time	Hospital ID	Chief Complaint	Patient Sex	Patient Age
11/28/2014	7:52:00	HOSP5	EVAUATION, DRANK COFFEE WITH CRUS	M	45-49
11/28/2014	7:53:00	HOSP5	DRANK TAINTED COFFEE	M	65-69
11/28/2014	7:57:00	HOSP5	DRANK TAINTED COFFEE	F	20-24
11/28/2014	7:59:00	HOSP5	INGESTED TAINTED COFFEE	M	35-39
11/28/2014	8:01:00	HOSP5	DRANK TAINTED COFFEE	M	45-49
11/28/2014	8:03:00	HOSP5	DRANK TAINTED COFFEE	M	40-44
11/28/2014	8:04:00	HOSP5	DRANK TAINTED COFFEE	M	30-34
11/28/2014	8:06:00	HOSP5	DRANK TAINTED COFFEE	M	35-39
11/28/2014	8:09:00	HOSP5	INGESTED TAINTED COFFEE	M	25-29

This detected cluster represents 9 patients complaining of ingesting tainted coffee, and demonstrates Semantic Scan's ability to detect rare and novel events.

First blinded user study

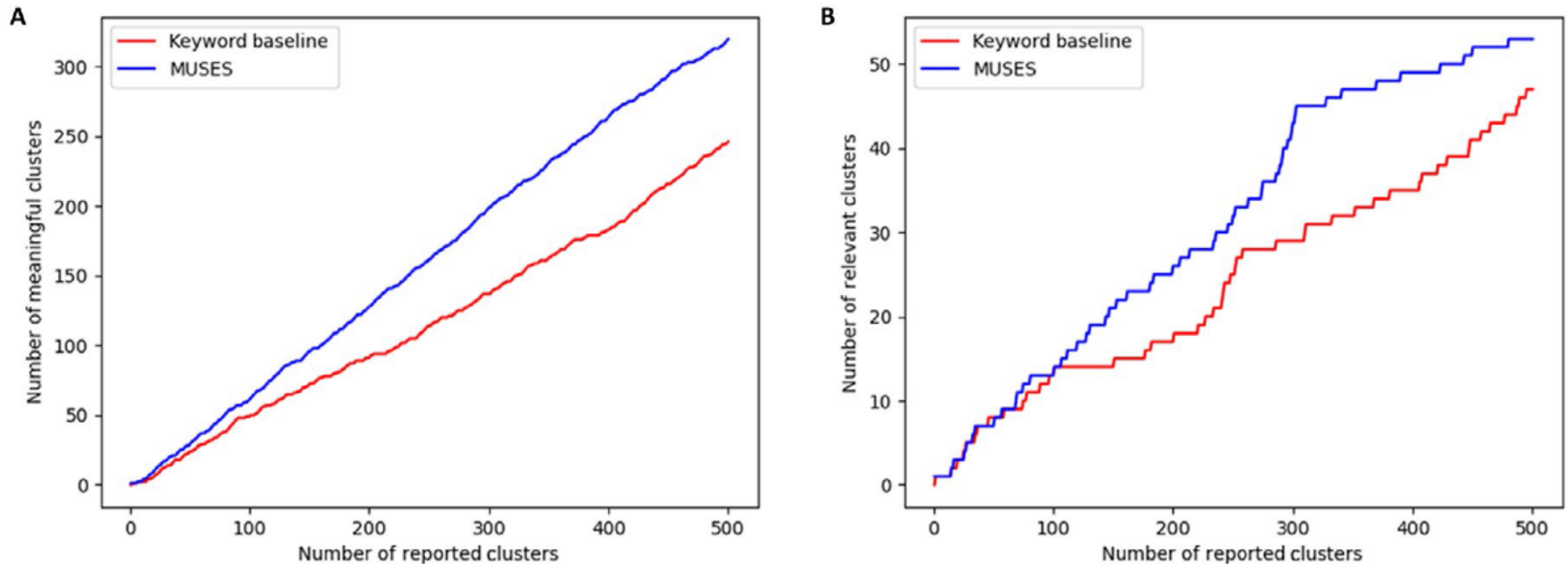
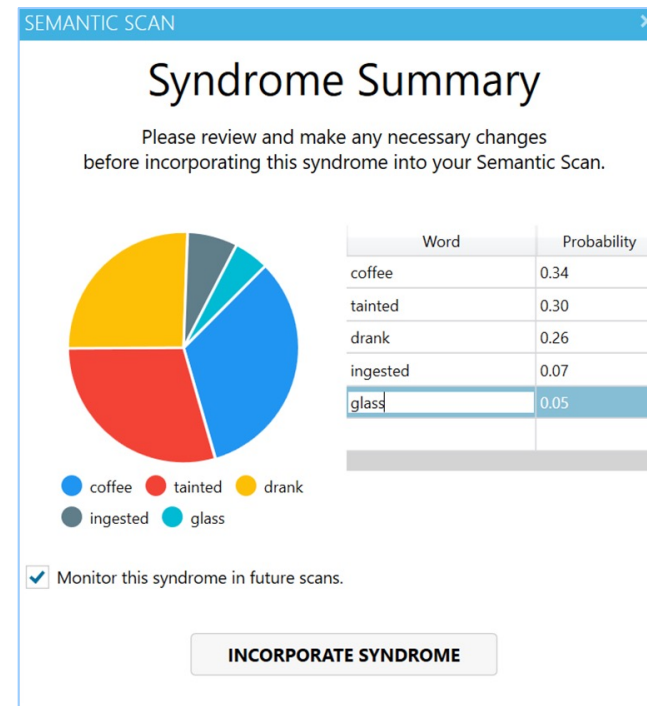
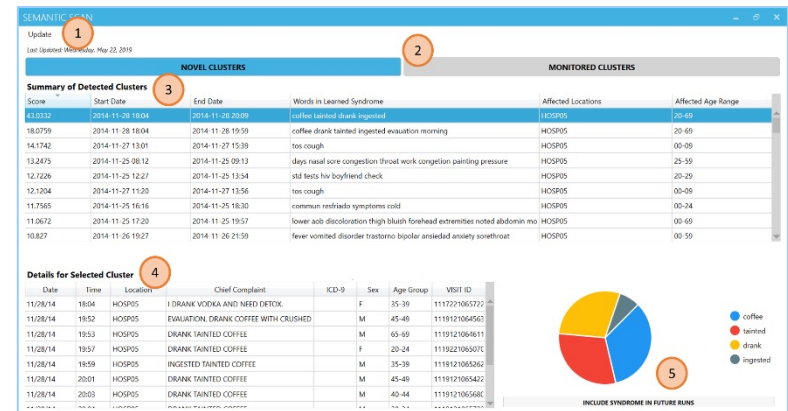


Fig. 2. Results from a blinded user study comparing our MUSES approach (fixed model) to a competing, keyword-based approach. Each method's top 500 highest-scoring clusters over a 6-year time period were rated as "meaningful and highly relevant," "meaningful but not highly relevant," or "not meaningful" by public health epidemiologists at NYC DOHMH. (A) Number of meaningful clusters and (B) number of "highly relevant" meaningful clusters, detected by each method, assuming that its top- k highest-scoring clusters were reported. Blue line: MUSES. Red line: keyword-based approach. For any fixed number of detected clusters, MUSES identifies more meaningful clusters and more highly relevant meaningful clusters than the keyword-based approach.

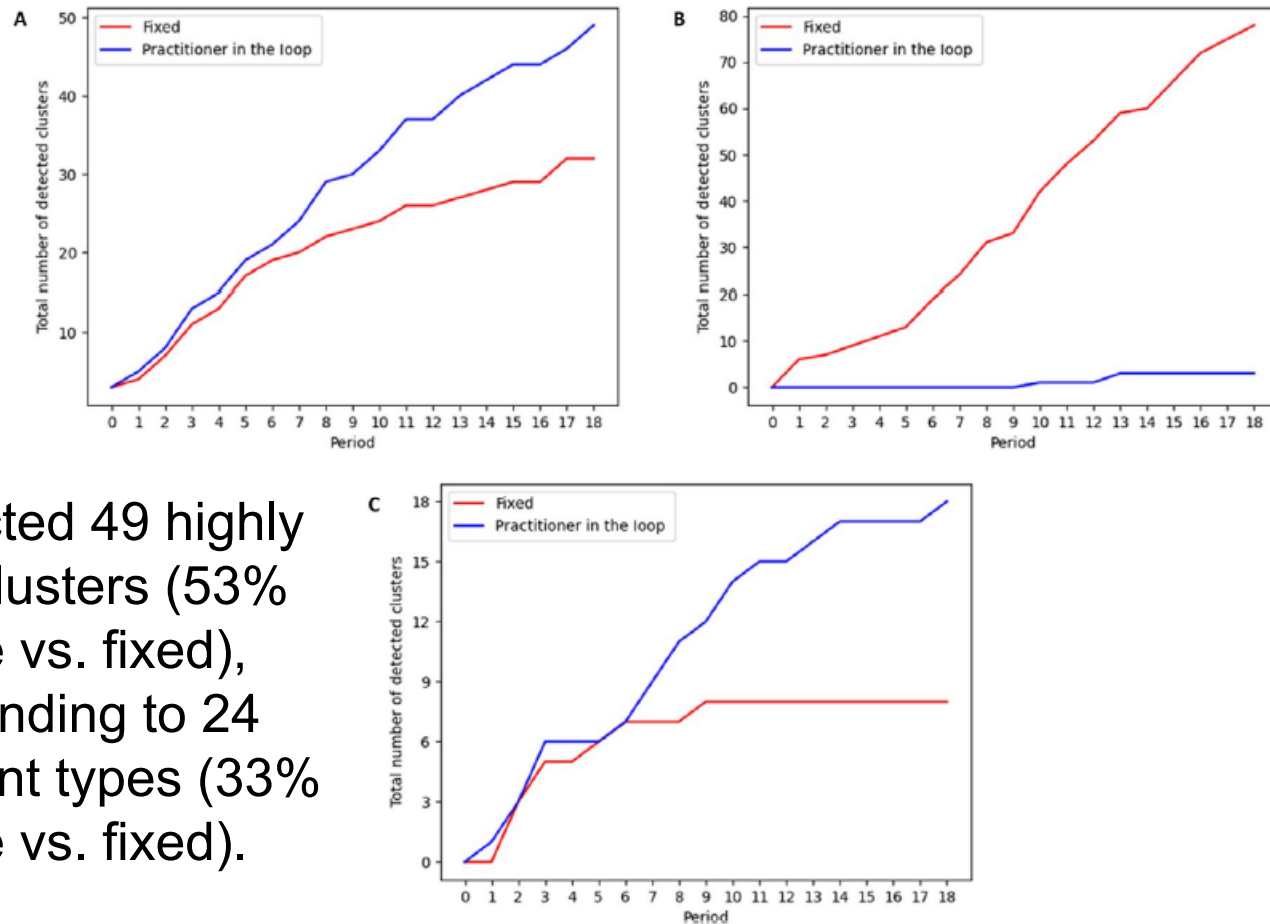
"Highly relevant" clusters included bacterial meningitis and synthetic drug use. "Meaningful" but not "highly relevant" clusters included motor vehicle accidents. "Not meaningful" clusters could be due to typos, coincidence, etc.

Incorporating user feedback

- Our system enables continual improvement of performance by including public health practitioners in the loop and incorporating their feedback.
- Users can add new syndromes and specify if they would like the system to monitor or ignore them in the future.
- Blinded user studies show that this Practitioner in the Loop approach enables the system to report more **relevant** clusters and to avoid overwhelming the user with irrelevant findings.



Second blinded user study



PITL detected 49 highly relevant clusters (53% increase vs. fixed), corresponding to 24 distinct event types (33% increase vs. fixed).

Fig. 4. Results from a blinded user study comparing the fixed and PITL models. Blue lines: PITL model. Red lines: fixed model. (A) Cumulative number of highly relevant clusters detected by each method, after each 2-week time period. The performance gap between the PITL and fixed models increases monotonically as a function of the number of labeled clusters used as training data by the PITL model. (B) Cumulative number of clusters detected by each method that were similar to clusters previously labeled "to ignore" by the user, after each 2-week time period. During the experiment, the fixed model detected 78 irrelevant clusters similar to those labeled "to ignore," while the PITL model only identified three such clusters. (C) Cumulative number of clusters detected by each method that were similar to clusters previously labeled "to monitor" by the user, after each 2-week time period. The PITL model identified a total of 18 highly relevant clusters that the practitioner had previously expressed interest in monitoring, as compared to 8 for the fixed model.

COVID results (March-June 2020)

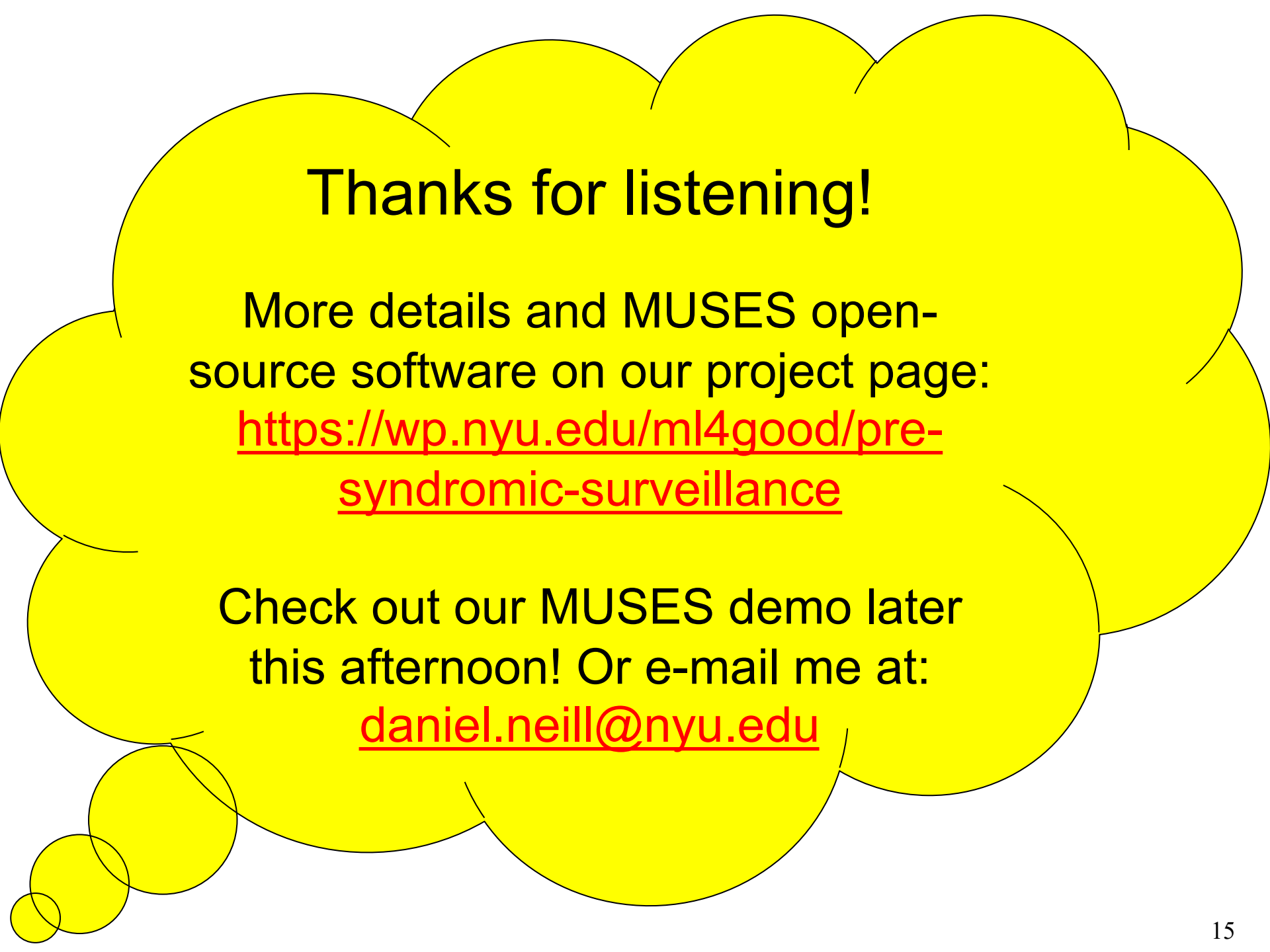
Table 2. Results from MUSES runs on ED chief complaint data from NYC DOHMH during the first wave of the novel coronavirus (COVID-19) pandemic in NYC, 1 March through 30 June 2020. Highest-scoring clusters found with 25 static and 25 emerging topics, scanning over both static and emerging topics. For each cluster, we report the date, de-identified hospital ID, number of cases, cluster duration in hours, whether the cluster is COVID-related, the most common chief complaints, and the cluster's log-likelihood ratio score. ICD-10 diagnosis codes were noted when used consistently to describe cases in the cluster (9 of 33 clusters). At least 30 of the 33 detected clusters were COVID-related. Thirty of 33 clusters occurred during the peak of the pandemic in NYC (17 March through 5 April), and 32 of 33 clusters corresponded to emerging topics rather than static topics.

Date	Hosp ID	No. of cases	No. of hours	COVID	Description	Score
27 March	31	164	12	Y	"Covid 19 exposure," flu-like symptoms, testing, cough, sob	244
28 March	31	152	10	Y	Testing, exposure, cough, sore throat, syncope	178
25 March	19	43	6	Y	"Coronavirus" [ICD-10: B97.29], cough, fever, headache, sob	75
29 March	31	111	11	Y	Testing, exposure, cough, fever, diarrhea, pneumonia	69
1 April	40	26	3	Y	Influenza-like respiratory [ICD-10: J10.1]	69
17 March	7	42	8	?	Smoke inhalation [ICD-10: J70.5], cough	65
26 March	1	14	3	Y	"Covid", cough, sore throat, body ache, measured O2	58
2 April	52	64	11	Y	screening for viral disease [ICD-10: Z11.59], cough, fever, sob	54
27 April	7	19	5	Y	"Covid 19 screening", cough, fever, sob	53
24 March	4	30	6	Y	Respiratory, headache	53

Discussion

Pre-syndromic surveillance is a **safety net** that can supplement existing ED syndromic surveillance systems by alerting public health to unusual or newly emerging threats.

Our recently proposed **multidimensional semantic scan** (MUSES) can accurately and automatically discover pre-syndromic case clusters corresponding to novel outbreaks and other patterns of interest.



Thanks for listening!

More details and MUSES open-source software on our project page:

<https://wp.nyu.edu/ml4good/pre-syndromic-surveillance>

Check out our MUSES demo later this afternoon! Or e-mail me at:

daniel.neill@nyu.edu



Michael Sheppard, MS
CDC



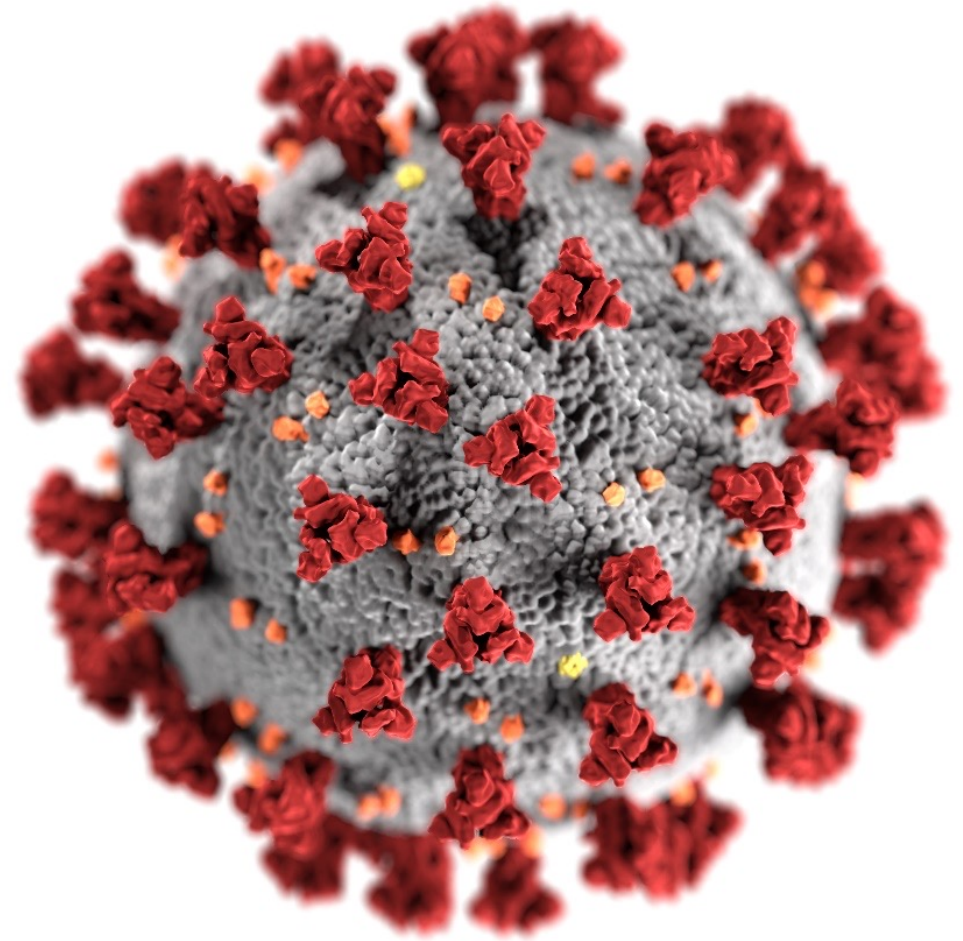
Natural Language Processing Methods for Assessing Changes in Emergency Department Symptom Presentation Among Patients Diagnosed with COVID-19 in the US, January 2021–August 2022

2022 Syndromic Surveillance Symposium

Michael Sheppard, MS

Centers for Disease Control and Prevention

Division of Health Informatics and Surveillance



cdc.gov/coronavirus

Natural Language Processing Task

- Obtain a large corpus of chief complaints associated with emergency department (ED) records with a diagnosis of COVID-19
 - Records identified using CDC COVID-specific diagnostic syndrome definition
- Use natural language processing (NLP) methods to identify symptom-based terms associated with each variant time period
 - Assess how symptoms have evolved
 - Assess disease severity during Omicron periods relative to pre-Delta and Delta periods

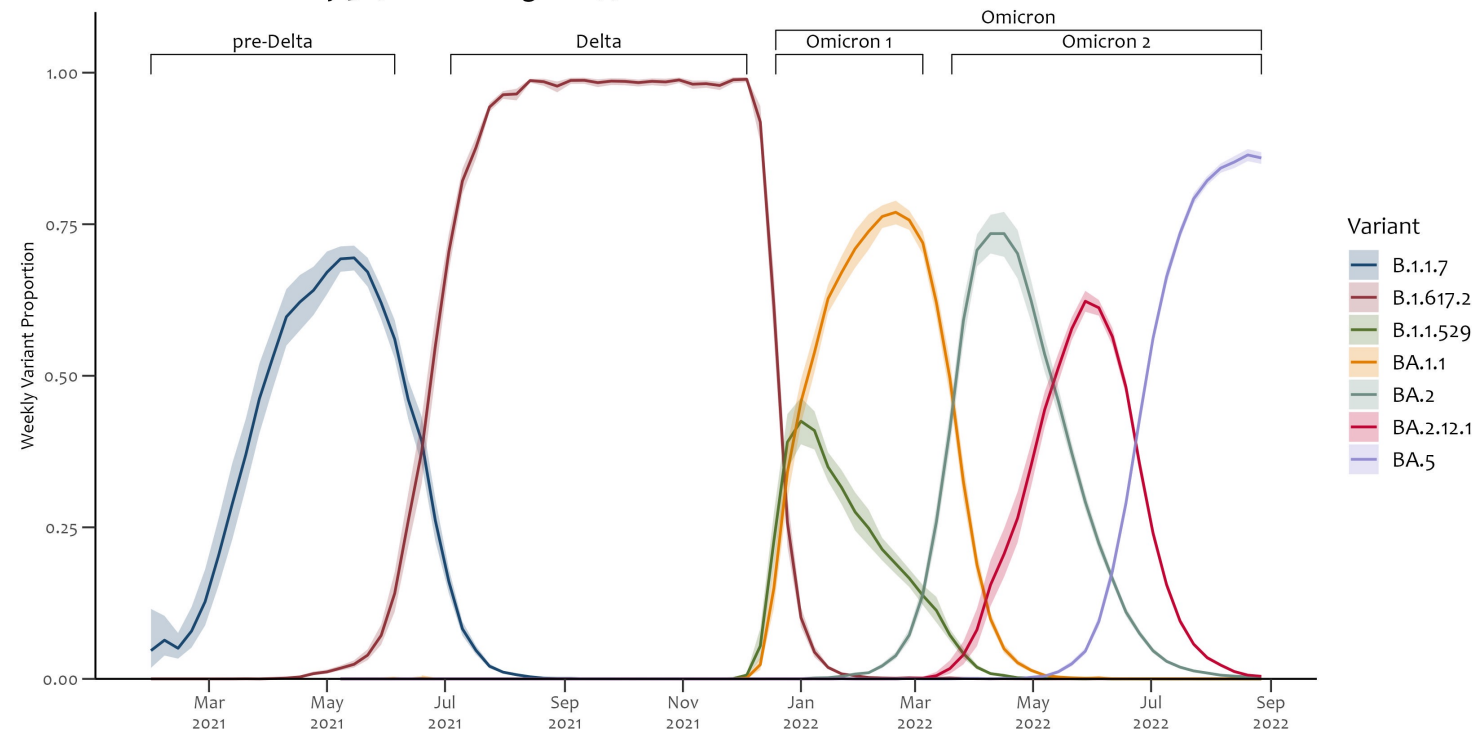
Using Genomic Surveillance to Determine Variant Time Periods

Variant proportions pulled from CDC COVID Data Tracker using Socrata APIs

Variant Time Period	Lineages Included	Date Range
Pre-Delta	1. B.1.1.7	December 27, 2020, to June 5, 2021
Delta	1. B.1.617.2	July 4, 2021, to December 4, 2021
Omicron 1	1. B.1.1.529 2. BA.1.1	December 19, 2021, to March 5, 2022
Omicron 2	1. BA.2 2. BA.2.12.1 3. BA.5	March 20, 2022, to August 27, 2022

Lineages exceeding 25% of national weekly COVID-19 infections between 2021 and 2022

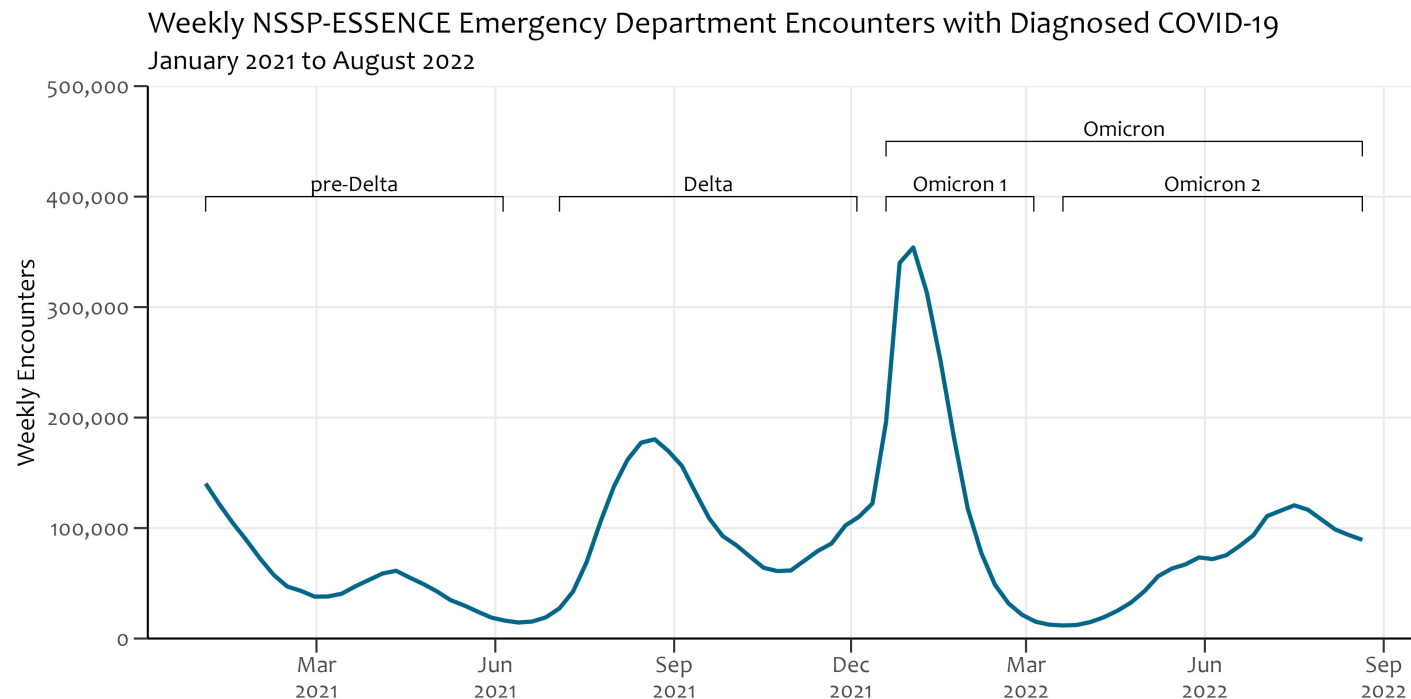
Weekly Proportion of Viral Lineages Among Infections for Major COVID-19 Variants in the USA: January 30, 2021 to August 27, 2022



Emergency Department Records from NSSP–ESSENCE with Diagnosed COVID-19

- 7,406,826 total chief complaints for ED visits with diagnosed COVID-19 pulled from week of December 27, 2020, to August 27, 2022
- Chief complaints grouped into variant time periods:

Variant Time Period	ED Encounters
Pre-Delta December 27, 2020, to June 5, 2021	1,307,254
Delta July 4, 2021, to December 4, 2021	2,244,268
Omicron 1 December 19, 2021, to March 5, 2022	1,932,928
Omicron 2 March 20, 2022, to August 27, 2022	1,597,045



Cleaning and Processing of Chief Complaint Field

- Chief complaints already *partially* processed by ESSENCE chief complaint parser
 - Removal of punctuation
 - Capitalization
 - Expansion of common chief complaint abbreviations

Original	Parsed
BREATHING DIFF	BREATHING DIFFICULTY
covid+, sob	COVID SHORTNESS OF BREATH
covid+, emesis;Emesis	COVID EMESIS EMESIS

Cleaning and Processing of Chief Complaint Field

- Additional standardization and cleansing of chief complaints were implemented to make the text suitable for NLP
 - Removing duplicate terms within the same chief complaint
 - Removing traditional and custom stop words
- Implemented in R by using prepackaged and custom stop word lists
- Terms removed *during* tokenization process

Cleaning and Processing of Chief Complaint Field

Term Removal Rule	ESSENCE-Parsed Chief Complaint	Final Chief Complaint
Remove discharge diagnosis codes from chief complaint	COUGH FEVER U071 HYPERTENSION I10	COUGH FEVER HYPERTENSION
Remove numeric terms	04 14 2022 NAUSEA VOMITING	NASUEA VOMITING
Remove standard English stop words	FEVER AND COUGH SHORTNESS OF BREATH	FEVER COUGH SHORTNESS BREATH
Remove common names that may appear in the chief complaint	TANYA FEVER	FEVER
Remove medical process terms and uninformative syndromic terms	BIBA AFTER TESTING COVID VOMITING SORE THROAT	COVID VOMITING SORE THROAT
Remove additional custom stop words that do not provide any symptom-based information	COUGH BODYACHES CONGESTION SINCE MONDAY	COUGH BODYACHES CONGESTION

Tokenization of Chief Complaints

- Tokenization: Splitting long strings of text into smaller units (tokens)
- n-gram: Contiguous sequence of n terms

Chief Complaint: SHORTNESS BREATH FEVER CHILLS TASTE SMELL COUGH

Tokens (bigrams, $n = 2$)



7,406,826 chief complaints tokenized in ***under 10 minutes*** using the R quanteda library

- R package for managing and analyzing textual data
- Supports basic and advanced NLP tasks
- Designed around infrastructures implemented in C++

<https://cran.r-project.org/web/packages/quanteda/index.html>

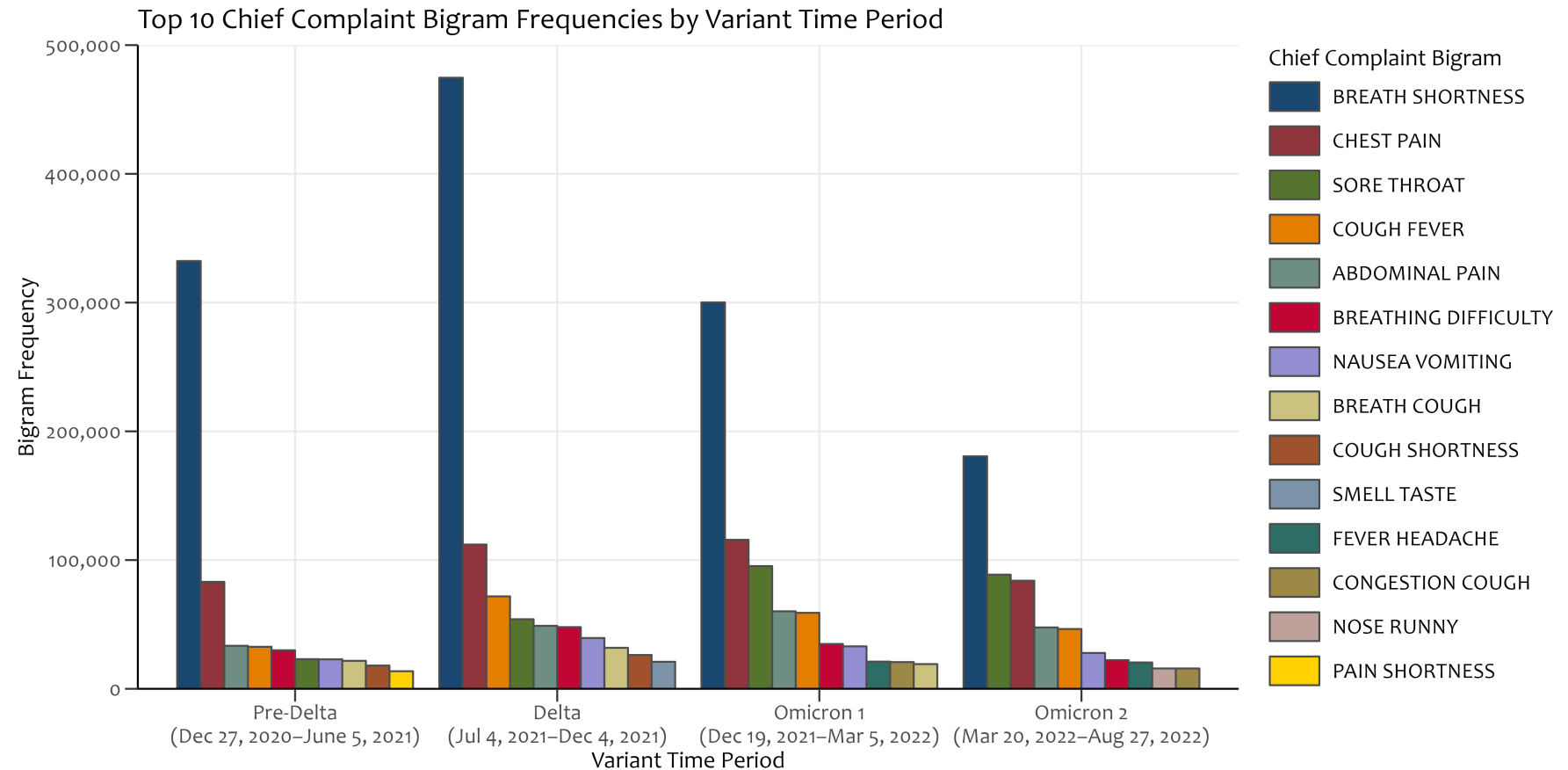
Traditional Methods for Summarizing Large Corpora of Text

- Find most common or characteristic symptoms of each variant time period by using simple methods for measuring term importance
 - Term frequencies
 - Term proportions
 - Term frequency-inverse document frequency (tf-idf)
- Bigrams (adjacent word pairs) are often more informative than individual terms
- Summarize the top n bigrams using one of the above metrics

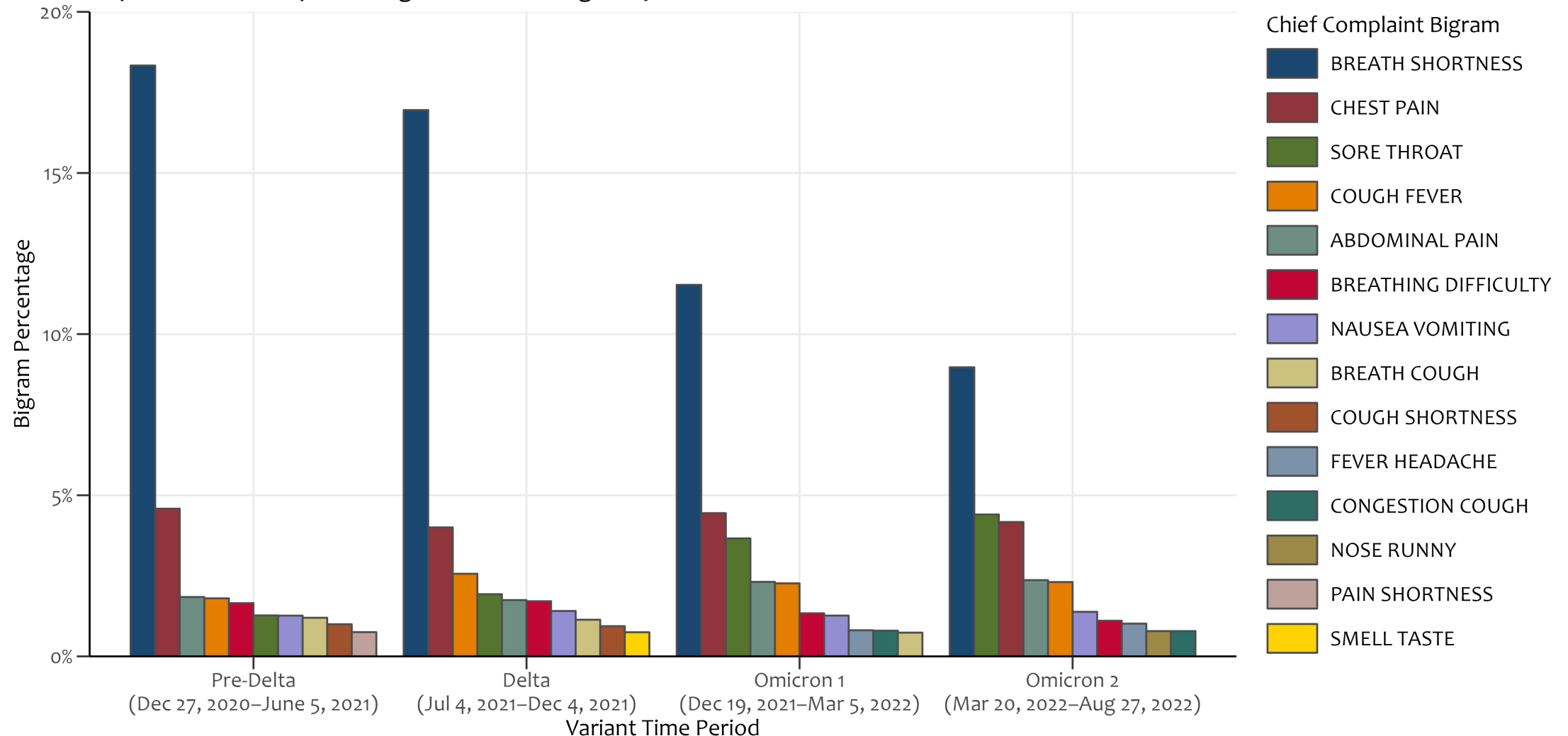
Method #1: Bigram Frequencies

Uninformative due to overlap in symptoms across all 4 periods

Only 14 unique bigrams across 4 periods



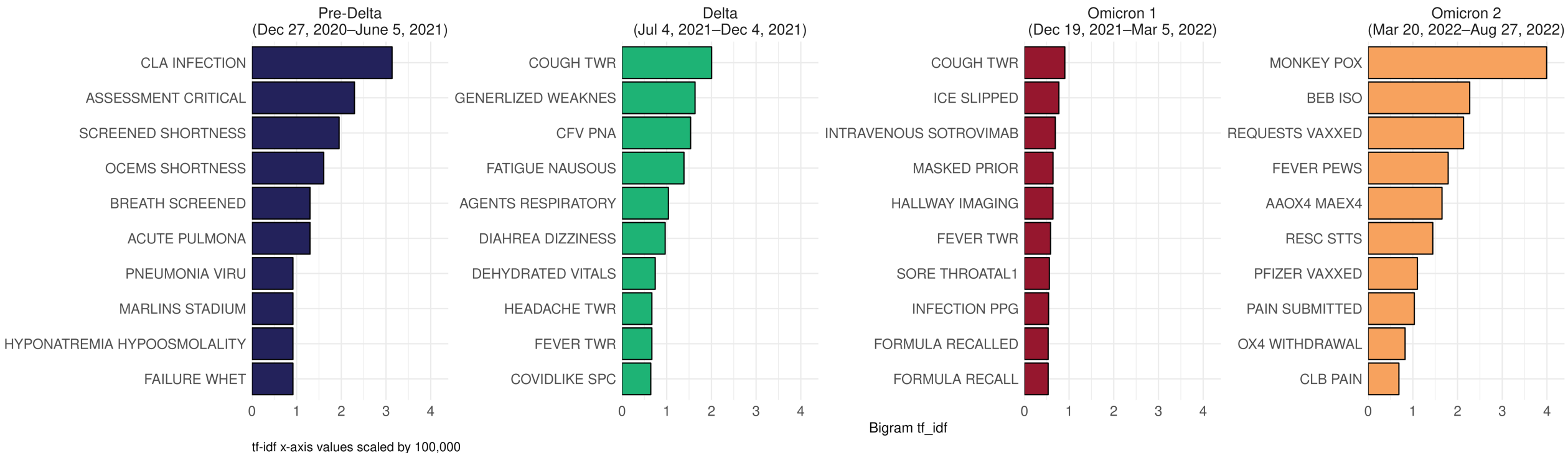
Top 10 Chief Complaint Bigram Percentages by Variant Time Period





Method #3: Term Frequency-inverse Document Frequency (tf-idf)

- Purpose: aims to identify the most important terms to a document, such as those which occur frequently in one but infrequently in others
- Disadvantage: tendency to identify **highly obscure** terms, especially for chief complaints and other free-text fields in syndromic data

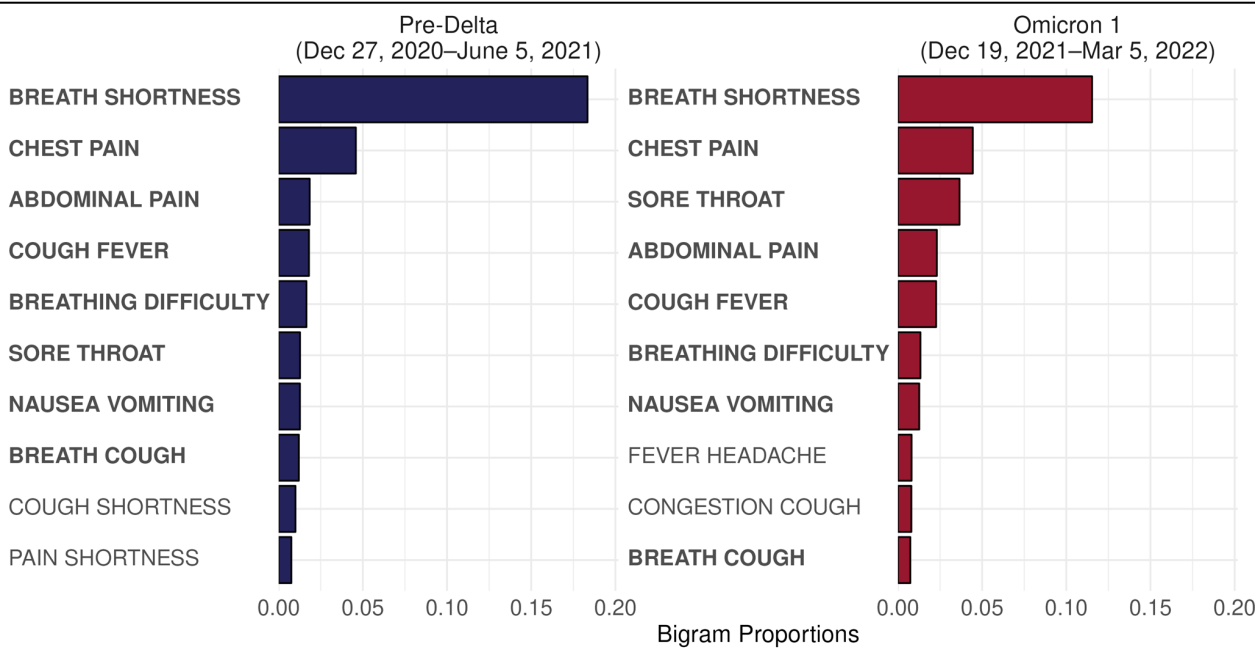


Problem: Traditional methods have disadvantages

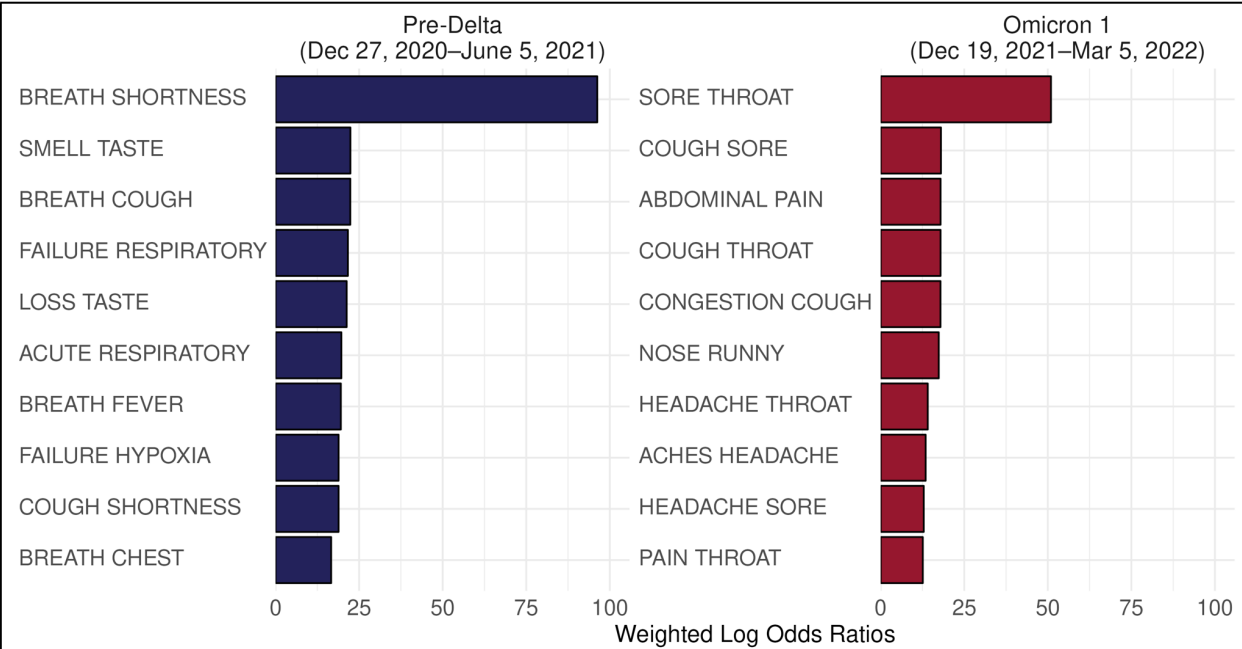
- Solution: Weighted log odds (WLO) ratio
 - Model-based approach that uses an **informative** prior
 - Identifies terms that are fundamentally important to a group, rather than those that are most frequent or obscure
 - Accounts for term variation across groups
 - Weights term estimates by their uncertainty (WLO ratios are z-scores)
- Implemented by the R package tidylo
 - <https://cran.r-project.org/web/packages/tidylo/tidylo.pdf>

Proportions vs. WLO Ratios for Identifying Symptoms Indicative of Variant Time Periods

Chief complaint bigrams with the 10 highest proportions for the Pre-Delta and Omicron 1 periods (**overlapping bigrams in bold**)



Chief complaint bigrams with the 10 highest WLO ratios[†] for the Pre-Delta and Omicron 1 periods (**no overlapping bigrams**)



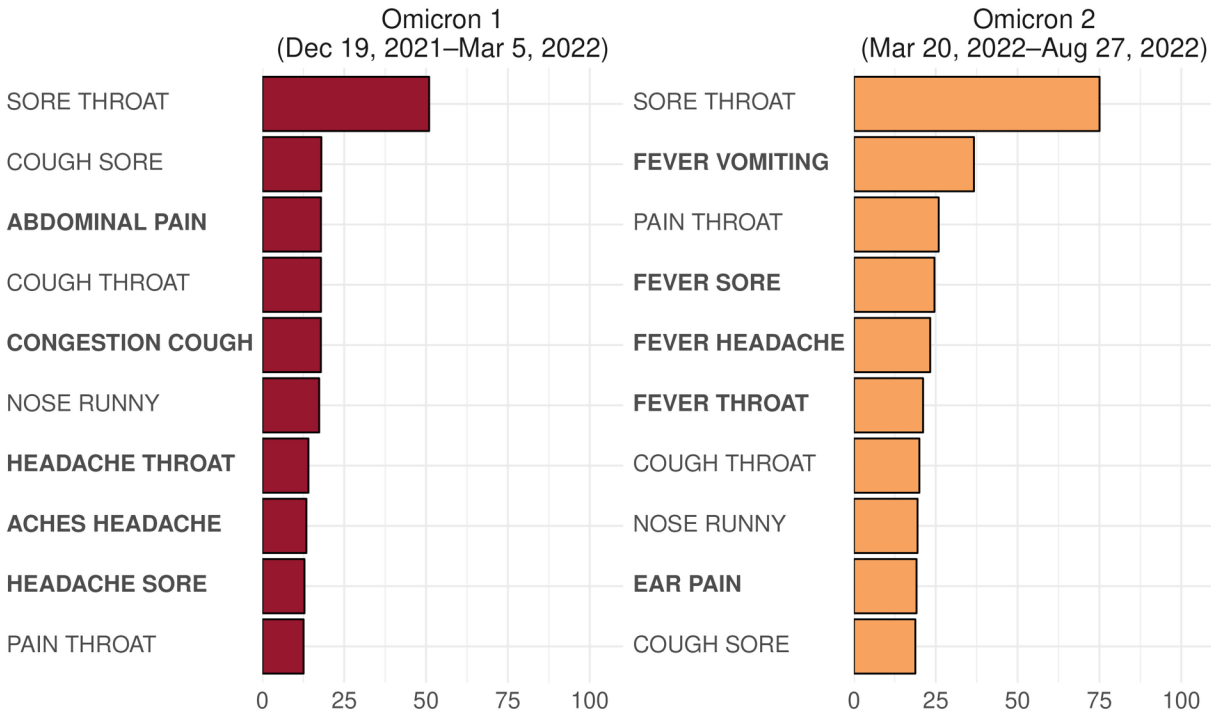
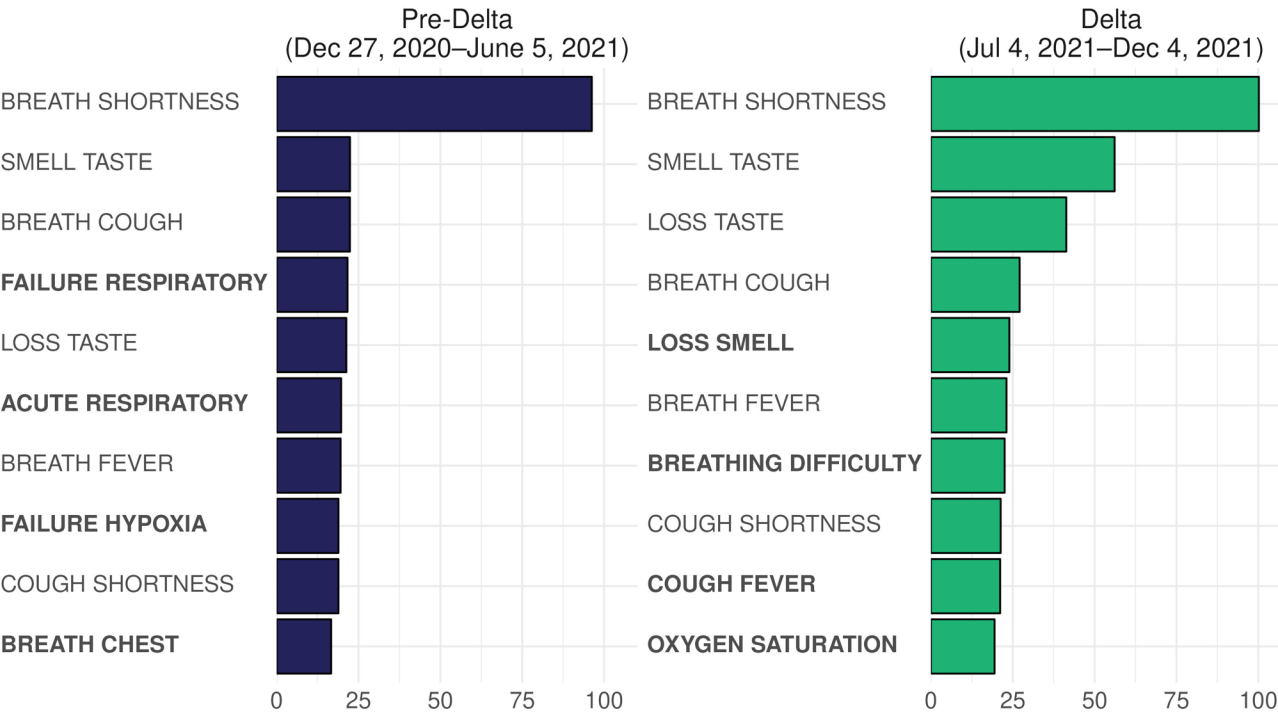
[†]Limited to Pre-Delta and Omicron 1 periods from WLO ratio comparison across all 4 variant time periods

Chief Complaint Bigrams with Top 10 Highest WLO Ratios by Variant Time Period

Symptoms perceived as being **more severe**

Symptoms perceived as being **less severe**

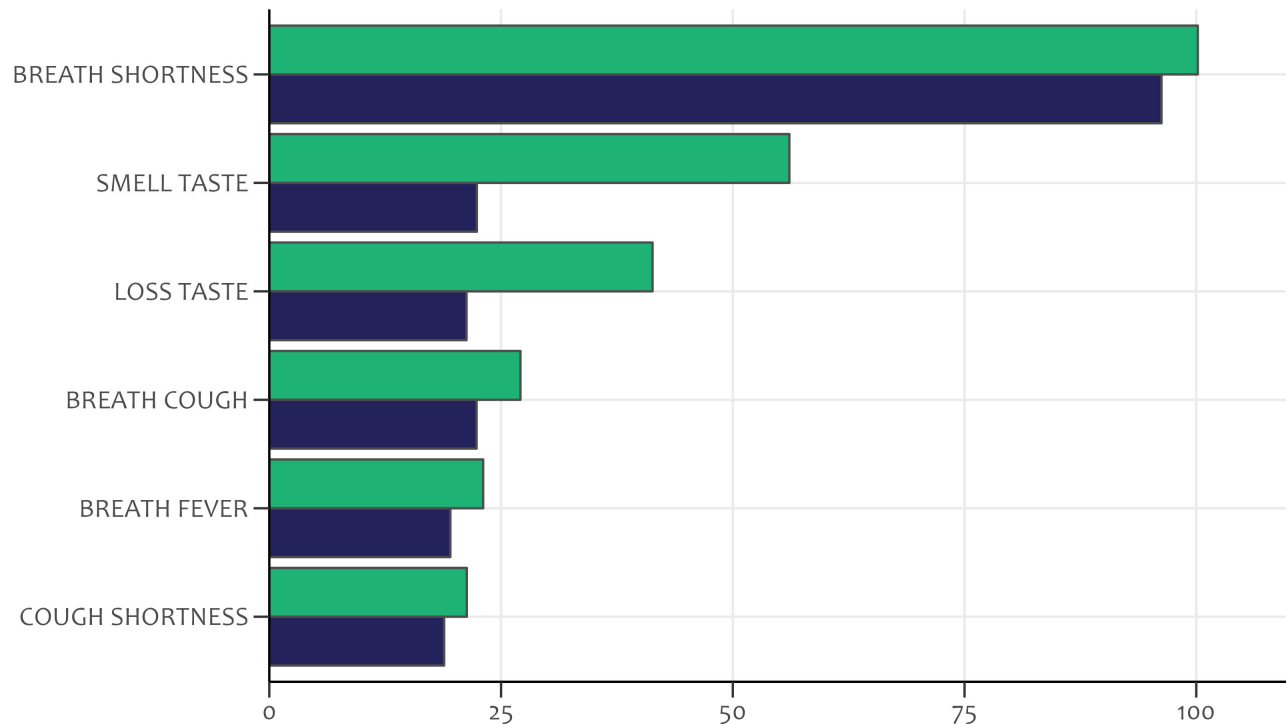
Bigrams Unique to One Period Presented in Bold



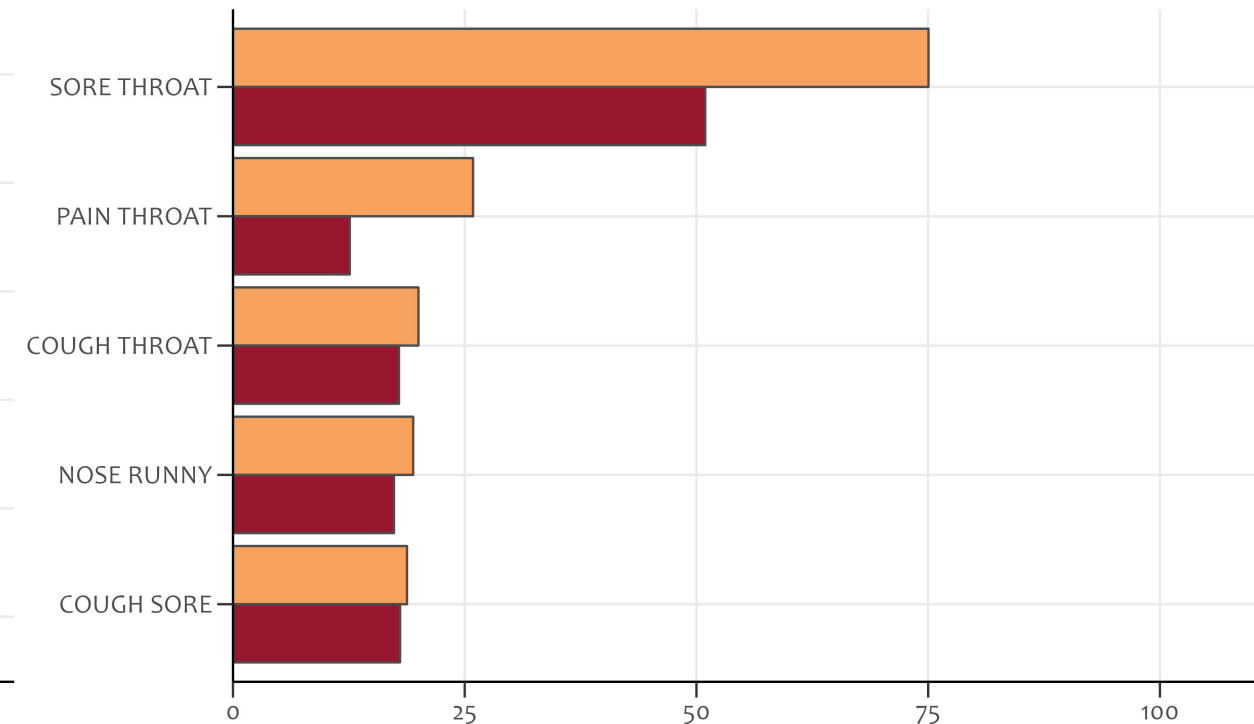
Chief Complaint Bigrams with Top 10 Highest WLO Ratios by Variant Time Period

Comparison of WLO Ratios for Bigrams Shared by Two Periods

Pre-Delta vs. Delta



Omicron 1 vs. Omicron 2

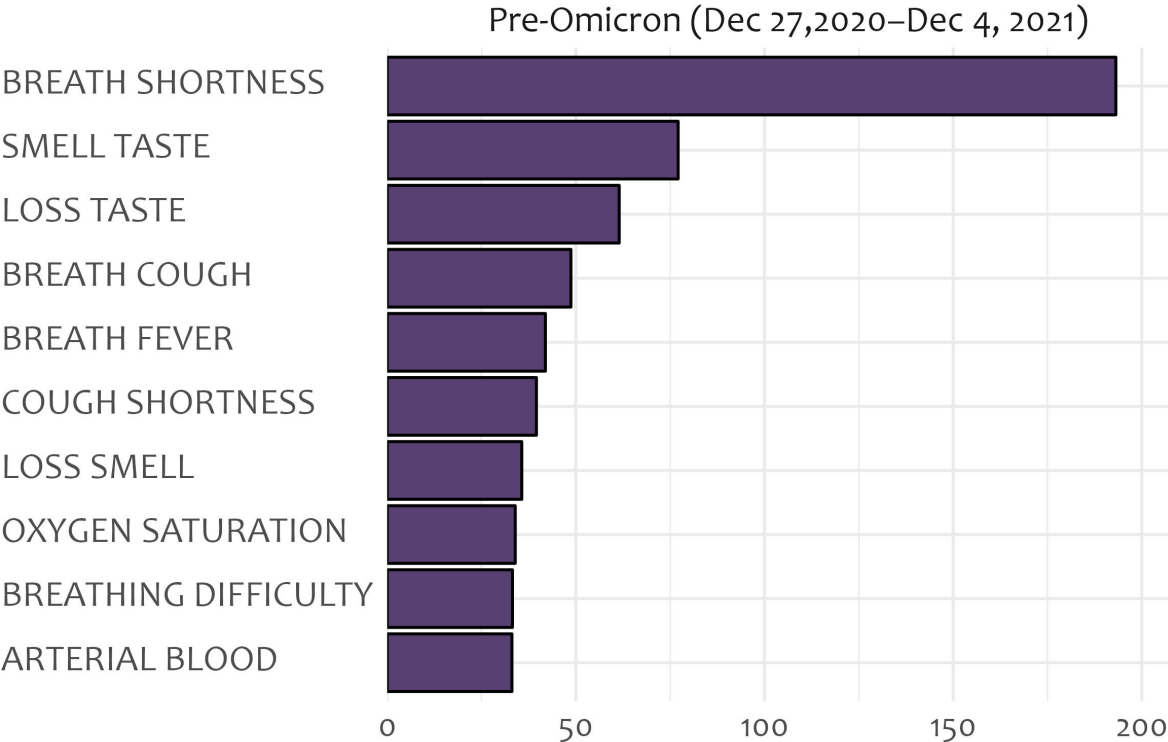


Weighted Log Odds Ratio

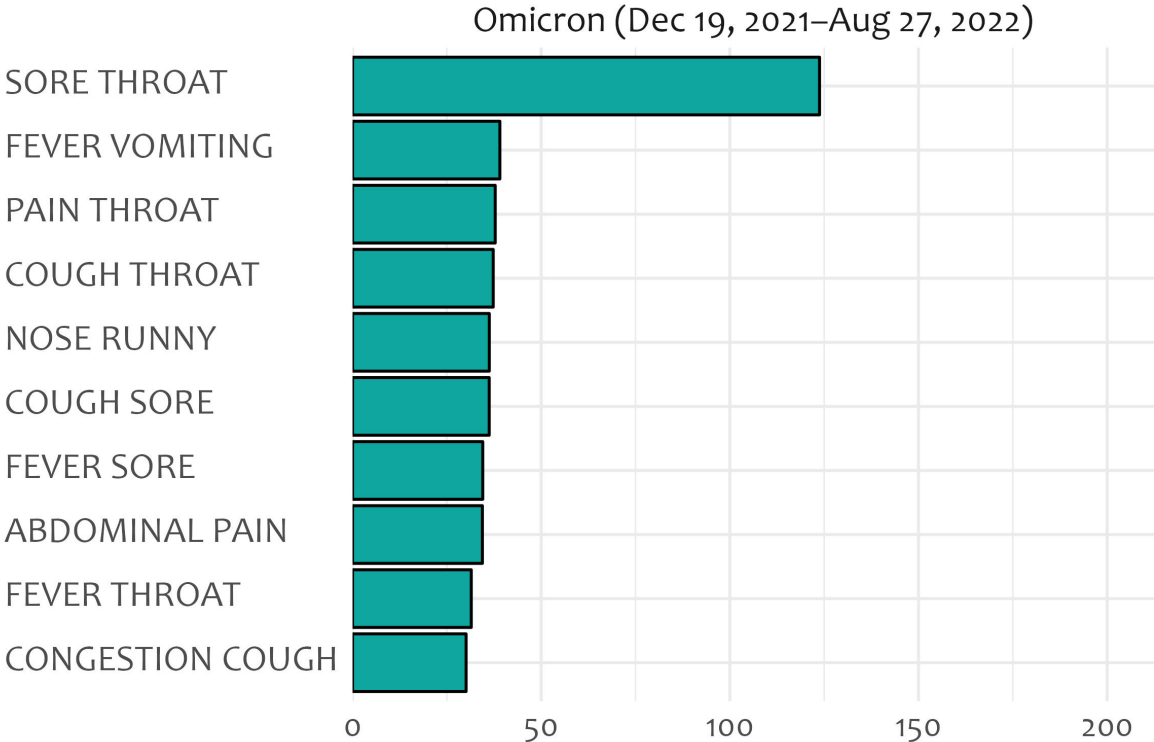


Alternative Comparison: Pre-Omicron vs. Omicron WLO Ratios

Grouping Pre-Delta and Delta together



Grouping Omicron 1 and 2 together

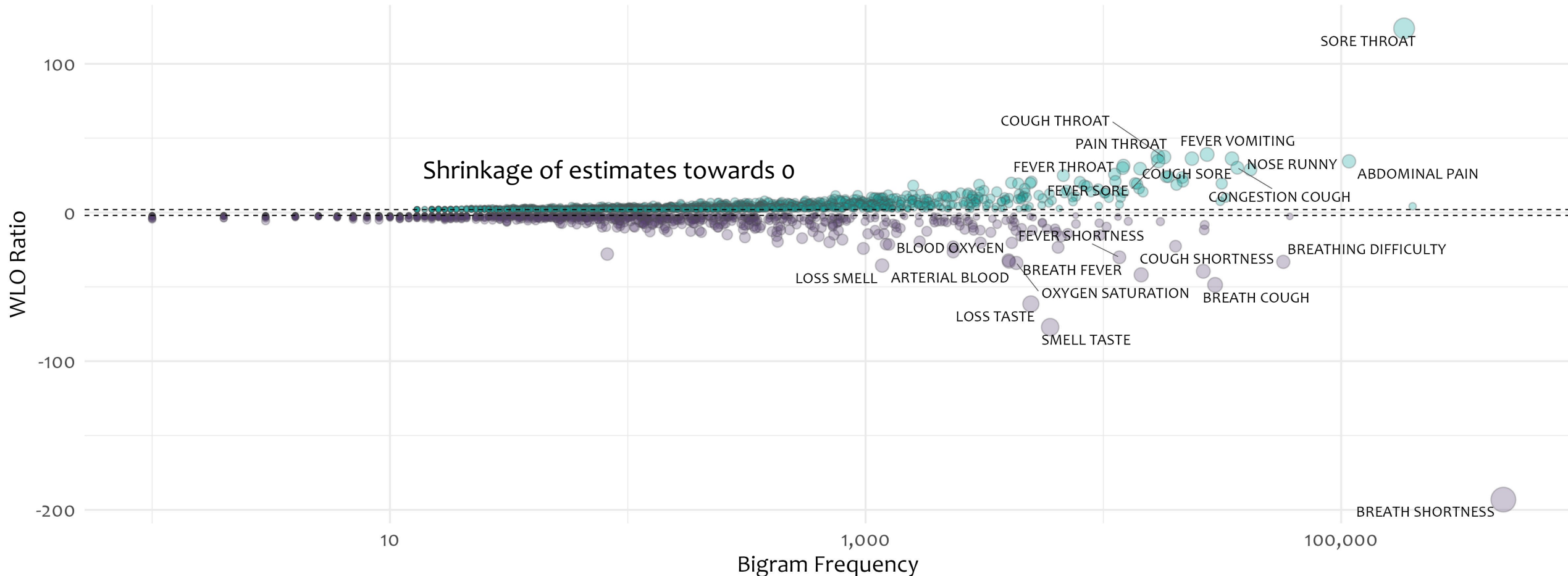


Weighted Log Odds Ratios

Bigram WLO Ratios vs. Frequencies for Pre-Omicron and Omicron Time Periods

WLO Ratios vs. Frequency for Omicron (Top) vs. Pre-Omicron (Bottom) Bigrams

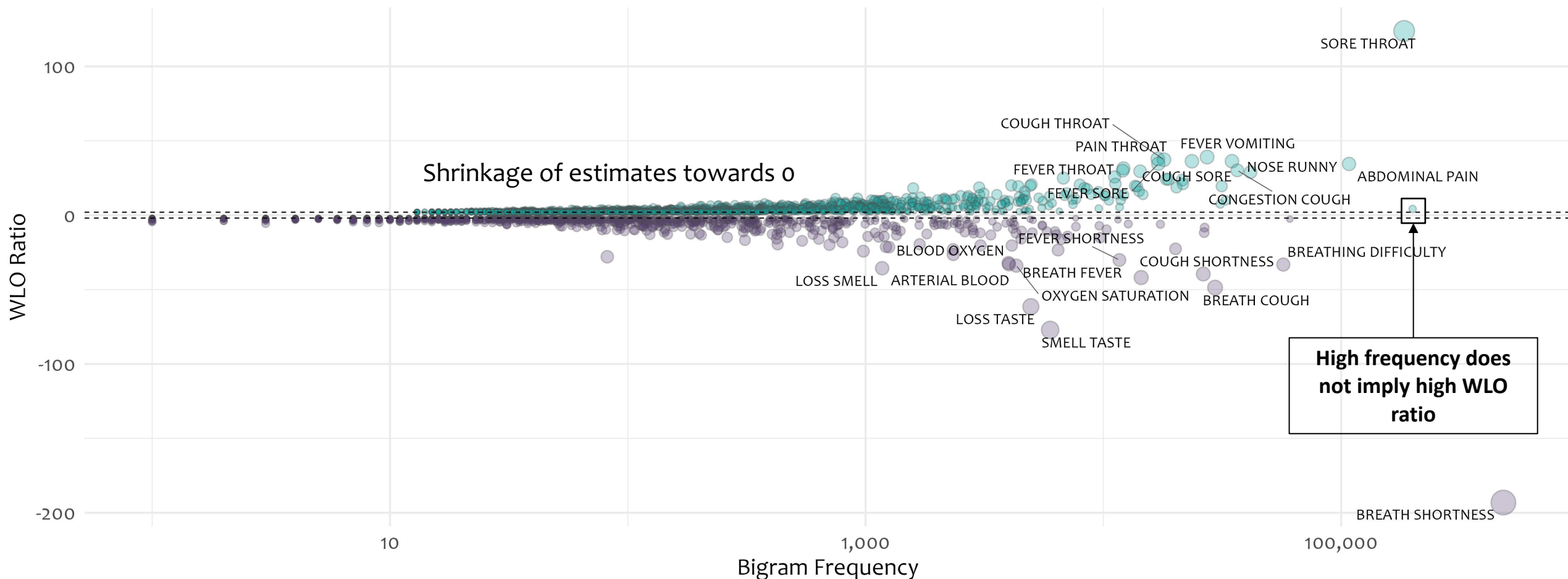
Limited to bigrams with absolute WLO ratio > 1.96



Bigram WLO Ratios vs. Frequencies for Pre-Omicron and Omicron Time Periods

WLO Ratios vs. Frequency for Omicron (Top) vs. Pre-Omicron (Bottom) Bigrams

Limited to bigrams with absolute WLO ratio > 1.96



Weekly Proportion Trends for Chief Complaint Bigrams with Top 10 WLO Ratios for Pre-Delta

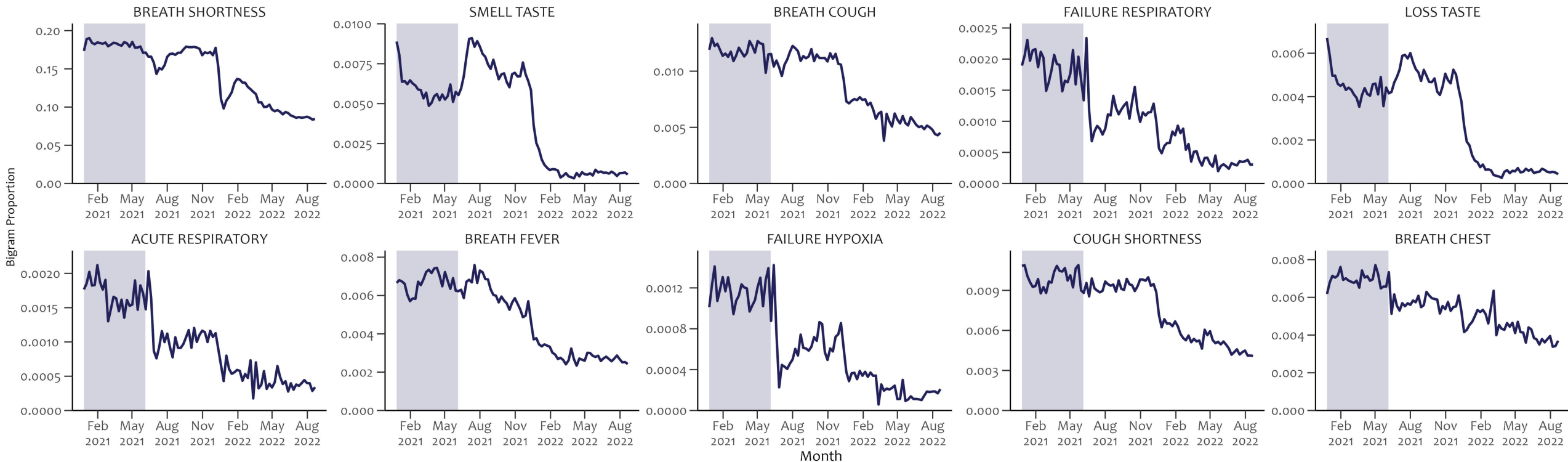
Example of weekly bigram proportion calculation:

$$\frac{y_{b_i,t}}{\sum_j y_{b_j,t}}$$

Frequency of bigram i in week t

Sum over all bigram frequencies occurring in week t

Weekly Chief Complaint Bigram Proportion Trends for pre-Delta Period



Weekly Proportion Trends for Chief Complaint Bigrams with Top 10 WLO Ratios for Delta Period

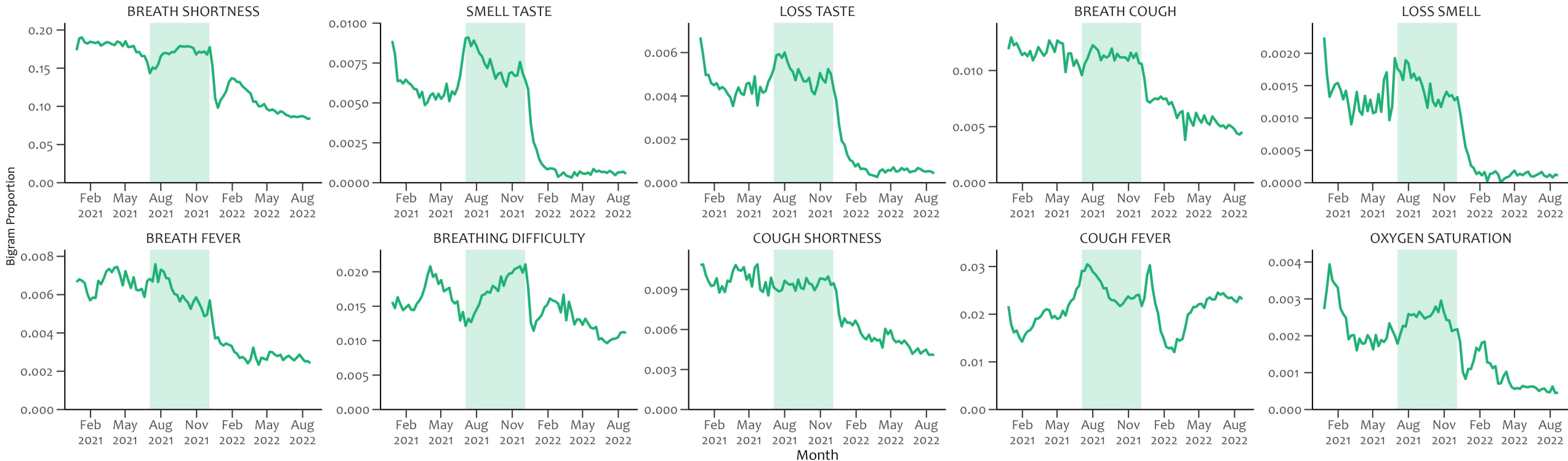
Example of weekly bigram proportion calculation:

$$\frac{y_{b_i,t}}{\sum_j y_{b_j,t}}$$

Frequency of bigram i in week t

Sum over all bigram frequencies occurring in week t

Weekly Chief Complaint Bigram Proportion Trends for Delta Period



Weekly Proportion Trends for Chief Complaint Bigrams with Top 10 WLO Ratios for Omicron 1 Period

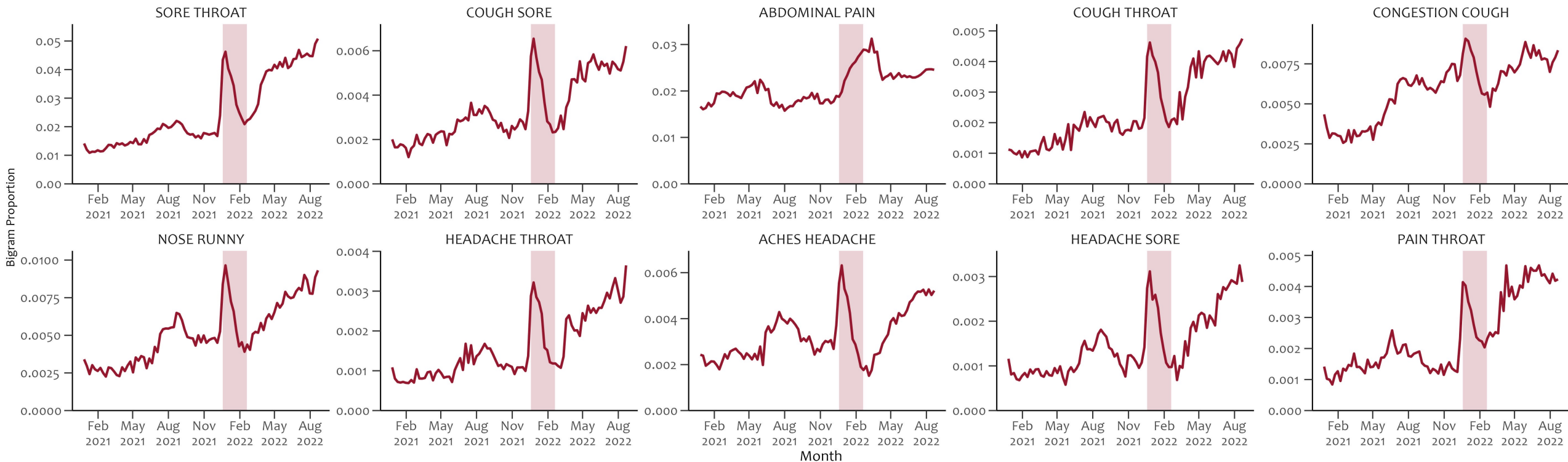
Example of weekly bigram proportion calculation:

$$\frac{y_{b_i,t}}{\sum_j y_{b_j,t}}$$

Frequency of bigram i in week t

Sum over all bigram frequencies occurring in week t

Weekly Chief Complaint Bigram Proportion Trends for Omicron 1 Period



Weekly Proportion Trends for Chief Complaint Bigrams with Top 10 WLO Ratios for Omicron 2 Period

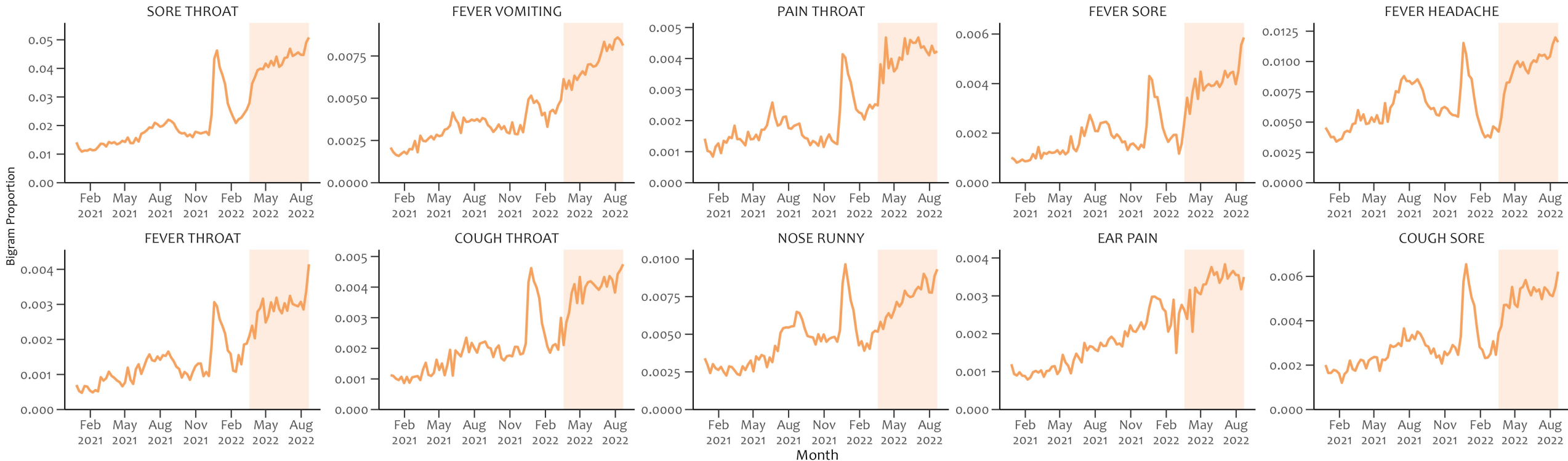
Example of weekly bigram proportion calculation:

$$\frac{y_{b_i,t}}{\sum_j y_{b_j,t}}$$

Frequency of bigram i in week t

Sum over all bigram frequencies occurring in week t

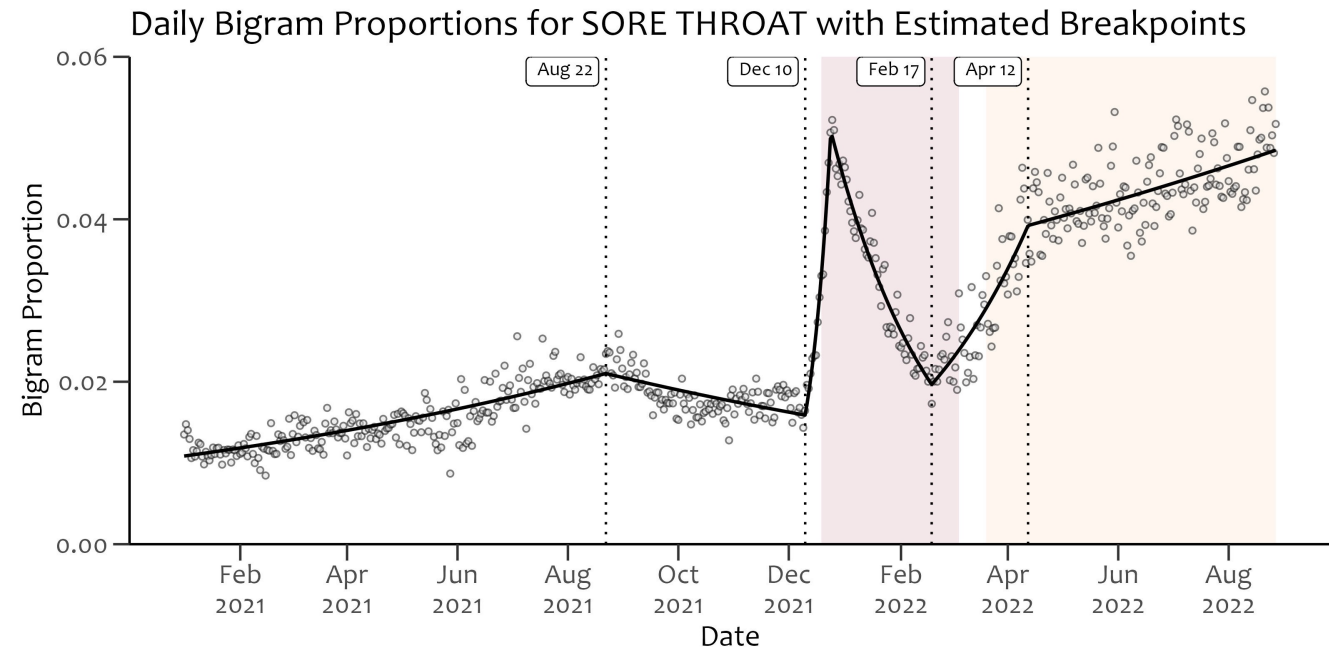
Weekly Chief Complaint Bigram Proportion Trends for Omicron 2 Period



Extensions of WLO Ratio Analysis – Breakpoint Estimation for Omicron Bigrams

- Use of segmented regression models to estimate breakpoints for which a change in slope occurred
 - Identify fundamental turning points in symptom presentation
- Useful for symptoms indicative of Omicron periods
 - Rapid shifts in symptom presentation

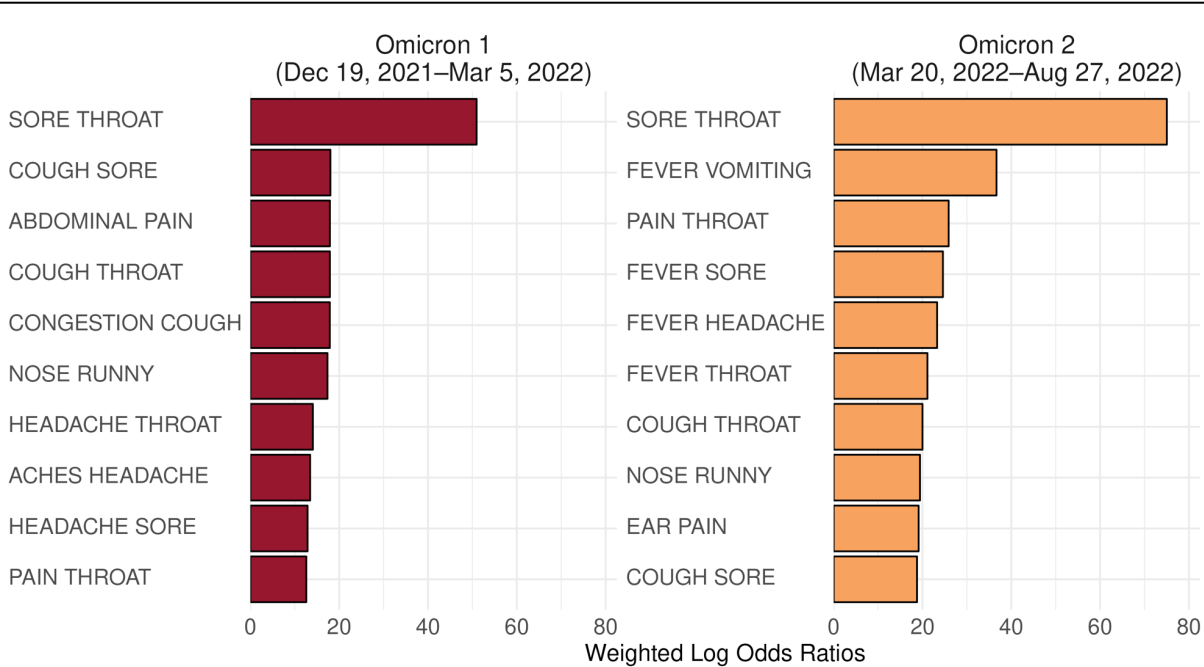
WLO Ratios for “SORE THROAT” During Omicron Variant Time Periods	
Variant Time Period	WLO Ratio
Omicron 1 December 19, 2021, to March 5, 2022	50.9
Omicron 2 March 20, 2022, to August 27, 2022	75.0



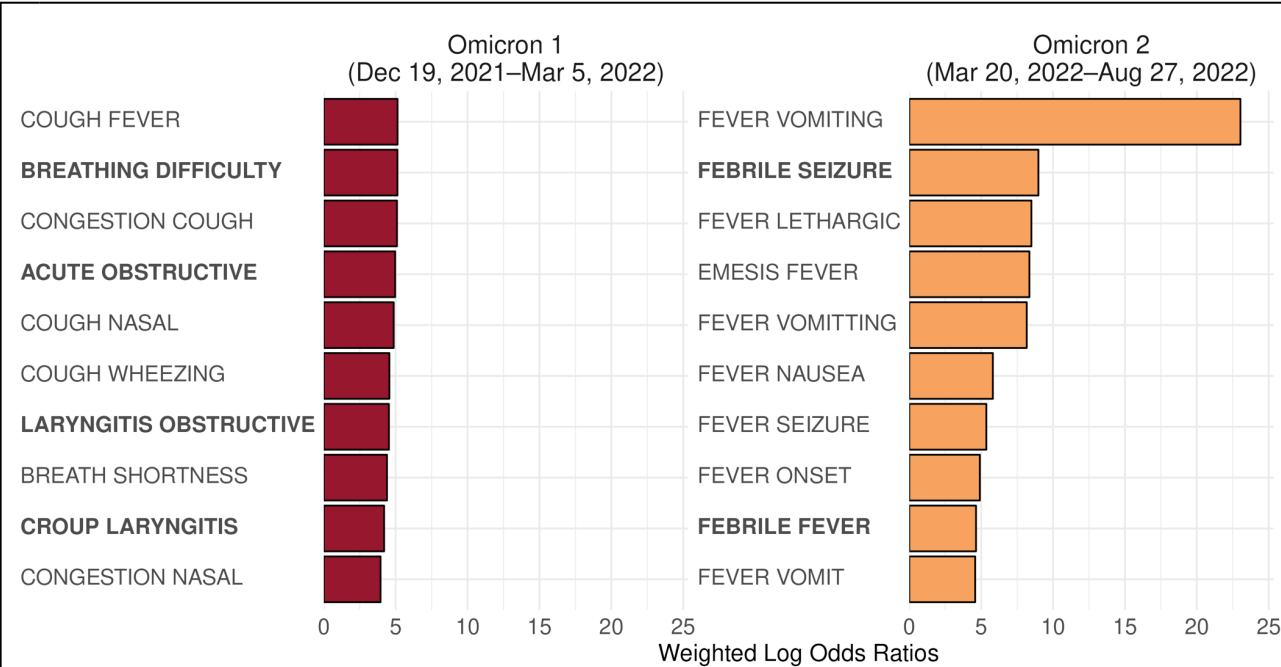
Top WLO Ratios for Omicron Variant Time Periods: All Ages vs. Children

- Limited to Omicron 1 and Omicron 2 periods from WLO ratio comparison across all 4 variant time periods
- Separate comparisons for all ages and 0–11 age group

WLO ratios for all ages



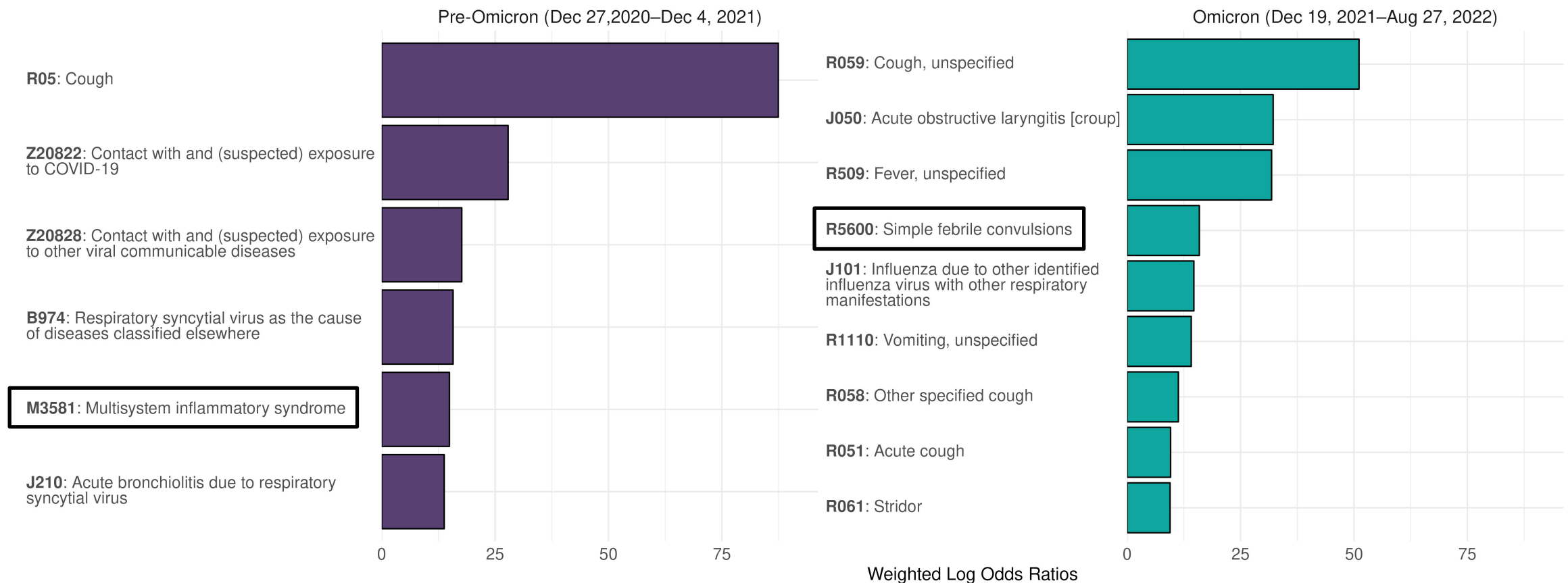
WLO ratios for children 0–11 with symptoms perceived to be more **severe** presented in **bold**



†Limited to Pre-Delta and Omicron 1 periods from WLO ratio comparison across all 4 variant time periods

Extensions of WLO Ratio Analysis – Applicable to Discharge Diagnoses and Other Fields

- Applied to discharge diagnosis codes for patients 0–11 years old
- Pre-Omicron vs. Omicron period



Conclusions

- WLO ratios can be used to identify symptoms that are most meaningfully representative of time periods aligning with predominance of major COVID-19 variants
 - Traditional methods have a disadvantage of identifying the same or most obscure terms in this context
- Findings suggest that relative to Pre-Delta and Delta periods, Omicron variants are characterized by less severe symptoms
 - Omicron 1: symptoms suggesting cold-like illness (congestion, cough, and sore throat)
 - Omicron 2: symptoms suggesting influenza-like illness (fever with vomiting, headache, and sore throat)

Conclusions

- Children younger than 12 years may still experience symptoms of Omicron variants that are perceived to be more severe
- WLO analyses are not restricted to free-text fields and can also be applied to:
 - Discharge diagnoses
 - Syndromes
- Bigram trend analyses can help identify turning points aligning with important shifts in symptom presentation

Limitations

- Perceived decreases in severity among patients during the Omicron period may be reflective of protective factors:
 - Higher rates of vaccination
 - Prior infection with COVID-19 during pre-Omicron periods
- Method does not distinguish between symptom descriptions that are clinically equivalent
 - “SORE THROAT” vs. “PAIN THROAT”
- Bigrams with the highest WLO ratios may be shared among periods when comparing 3 or more categories

Acknowledgements

Centers for Disease Control and Prevention

Aaron Kite-Powell

Kathleen Hartnett

Karl Soetebier

Jennifer Adjemian

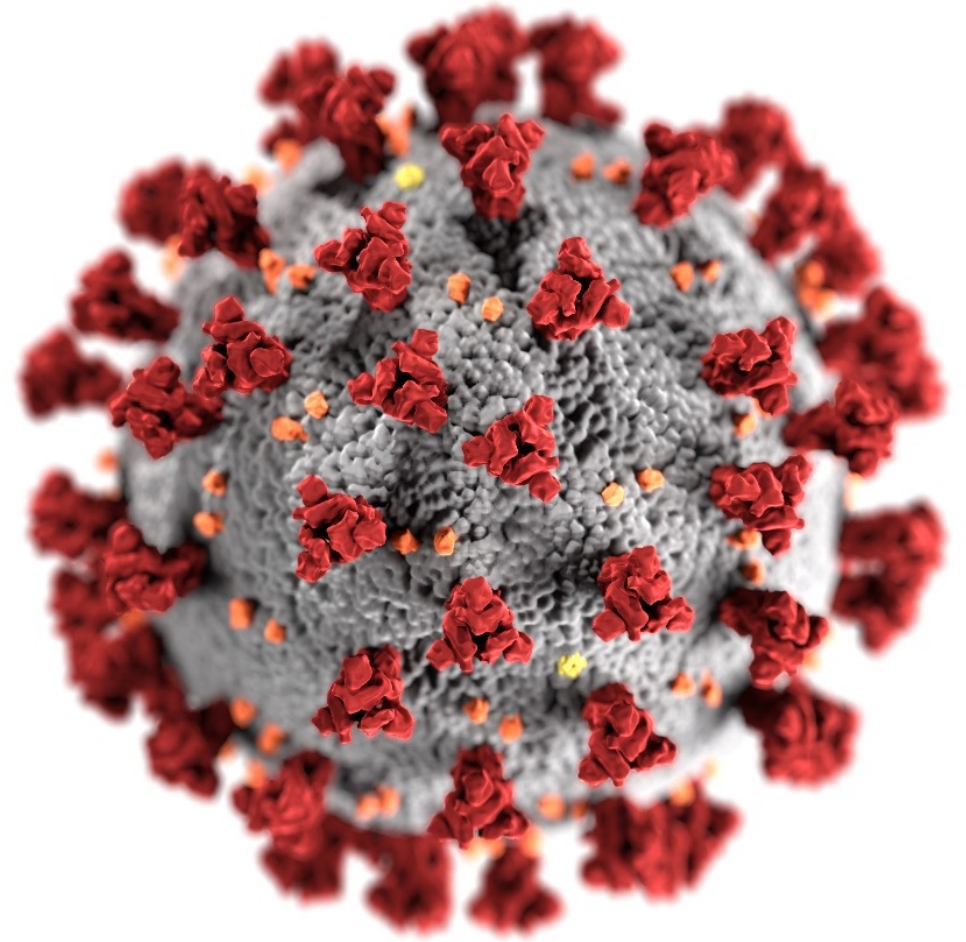
Thank You!

Contact Information

Please feel free to contact me at OUL2@cdc.gov with any additional questions!

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

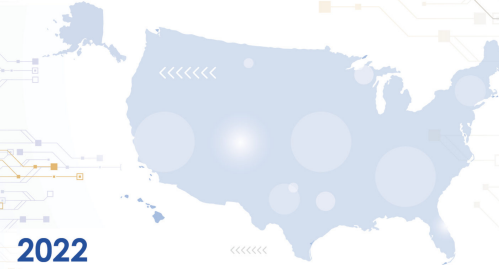
The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.



National Syndromic Surveillance Program

Community of Practice

Resources that Advance the Science and
Practice of Syndromic Surveillance



2022

Syndromic Surveillance Symposium

December 6–8, 2022



KAHUIUNA
CONSULTING