COVID-19 Impacts on Primary Care Practice Patterns
A Machine Learning Evaluation

Christopher Meaney

Biostatistician
Department of Family and Community Medicine
University of Toronto
&
PhD Candidate
Division of Biostatistics
Dalla Lana School of Public Health
University of Toronto

May 20, 2021

Acknowledgements

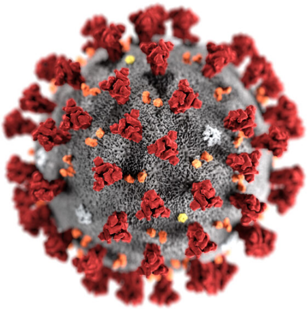
- UTOPIAN: Michelle, Karen, Braden, Sumeet, Babak, Tao, Rabiya, many others...
- DFCM: Rahim, Peter, Eva, Paul, Mary-Ann, Bojana, Julia, Teja, many others...
- ICES: Liisa, Therese, Peter.
- DLSPH: Mike (PhD Supervisor) & entire biostats program.
- NYFHT: Patients and physicians for generating the primary care progress note corpus.

Section 1: Background

- Defining COVID-19.
- Measuring the impact of COVID-19 on Toronto, Canada.
- Metrics for evaluating the impact of COVID-19 on community and public health.
- Comparison of traditional vs. text-based designs for monitoring/evaluating COVID-19.
- Specification of study research questions.

What is COVID-19?

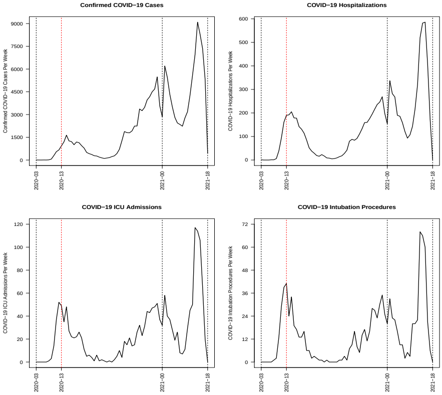
Scope Note A viral disorder generally characterized by high FEVER; COUGH; DYSPNEA; CHILLS; PERSISTENT TREMOR; MUSCLE PAIN; HEADACHE; SORE THROAT; a new loss of taste and/or smell (see AGEUSIA and ANOSMIA) and other symptoms of a VIRAL PNEUMONIA. In severe cases, a myriad of coagulopathy associated symptoms often correlating with COVID-19 severity is seen (e.g., BLOOD COAGULATION; THROMBOSIS; ACUTE RESPIRATORY DISTRESS SYNDROME; SEIZURES; HEART ATTACK; STROKE; multiple CEREBRAL INFARCTIONS; KIDNEY FAILURE; catastrophic ANTIPHOSPHOLIPID ANTIBODY SYNDROME and/or DISSEMINATED INTRAVASCULAR COAGULATION). In younger patients, rare inflammatory syndromes are sometimes associated with COVID-19 (e.g., atypical KAWASAKI SYNDROME; TOXIC SHOCK SYNDROME; pediatric multisystem inflammatory disease; and CYTOKINE STORM SYNDROME). A coronavirus, SARS-CoV-2, in the genus BETACORONAVIRUS is the causative agent.

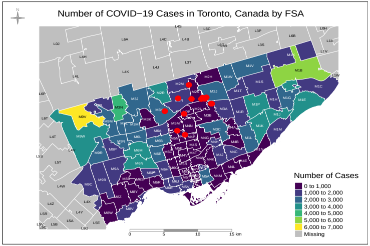


National Library of Medicine (Medical Subject Headings): COVID-19.

Monitoring the Impact of COVID-19 in Toronto, Canada

- Three distinct waves of COVID-19 infection.
- 152,215 lab confirmed COVID-19 infections (as of May 2021).
- 9338 ever-hospitalized; 1714 ever-ICU; 985 ever-intubated.





Open Data Toronto: COVID-19 Dataset and Toronto FSA Shape File (Retrieved May 2021).

Metrics Quantifying Direct Effects of COVID-19 on Morbidity/Mortality

- Laboratory confirmed COVID-19 cases.
- COVID-19 related hospitalizations.
- COVID-19 related ICU admissions.
- COVID-19 ICU patients who are medically intubated.
- COVID-19 related deaths.
- COVID-19 vaccine doses administered. Numbers with 1, 2, etc. vaccinations.

Myriad of Indirect Mechanisms COVID-19 Impacts Community Health

- Reduced/altered access to primary healthcare (family medicine, dentistry, pharmacy, etc.).
- Limited access to referral-based specialist care (secondary/tertiary-care).
- Delayed/cancelled elective medical procedures (e.g. surgeries, imaging, etc.).
- Postponed screening/preventative-care (e.g. childhood vaccinations, cancer screens, etc).
- Decreased monitoring/recording of routinely collected clinical measures (e.g vitals, labs).
- Altered health behaviours (e.g. smoking, drinking, food intake, exercise, sleep, etc.).
- Changes in health determinants and disease burden (incidence, prevalence, severity).

Common Elements of COVID-19 Primary Care Research Studies

- Research Questions: "Impact of COVID-19 on XXX in patient group YYY in setting ZZZ".
- Design design is typically observational/quasi-experimental.
- Multitude possible data sources: admin-, registry-, survey-, cohort-, EMR-data.
- Outcome measures potentially complex operationalization, constructed from observed data.
  - "Validating" construct/phenotype/identification-algorithm is challenging.
- Given an operationalized metric - explore variation over time (in response COVID-19).

Monitoring/Evaluating COVID-19 Using a Text-as-Outcome Design

- Research Questions: "Impact of COVID-19 on PC practice patterns in Toronto, Canada".
- Study Design: retrospective open cohort.
- Data Source: EMR text data (unique, rich/expressive, technical, diverse, reliable).
- Outcome: unsupervised thematic/topical phenotype of primary care (learned from text).
- Methods: monitor evolution of latent primary care topics (over time) in response COVID-19.

Benchimol et al (2015). REPORT Statement. PLoS Medicine.

Rosella et al (2010). Importance Accurately Identifying Disease in EHR Studies. BMJ.

Research Objectives: Using a large collection of primary care progress notes from Toronto, Canada obtained between 01/01/2017 and 31/12/2020:

- Estimate a meaningful topical/thematic basis characterizing primary care practice patterns.
- Monitor and evaluate how practice patterns (topical bases) evolve over time.
- Identify how the COVID-19 impacts primary care practice patterns and community health.

Research in Context

- Primary care text data contains unique information (not available in other data sources).
- Unsupervised monitoring of COVID-19 using passively collected data is cost-effective tech.
- Inferences from unsupervised evaluation of text data complement traditional methods.
- Holistic understanding COVID-19 impacts useful for post-pandemic planning/prioritization.

Section 2: Design, Text Processing & Corpus Characteristics

- Design, Setting, Measures and Inclusion/Exclusion Criteria.
- Computationally Processing Text Data into a Document Term Matrix.
- Description of Sample/Corpus Characteristics.

Design: Retrospective open cohort.

Setting: North York Family Health Team

- 12 clinical sites geographically distributed across North-Central Toronto, Canada.
- Single shared/integrated electronic medical record (EMR) system.

Data Source: UTOPIAN Q4-2020 database.

- Clinical notes table. Patient table.

EMR Measures:

- Note ID, Patient ID, Physician ID, Site ID.
- Patient Age, Sex, FSA.
- Progress Note Text, Date Progress Note Recorded.

Inclusion/Exclusion Criteria: (Note-Level Unit-of-Analysis)

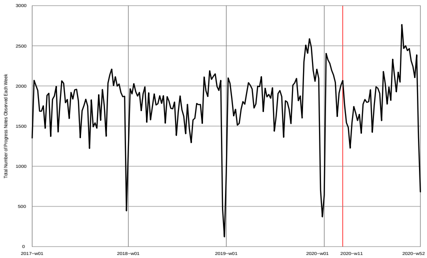
- Include: 418,223 NYFHT progress notes from 01/01/2017-31/12/2020.
- Exclude: 80 notes missing age.
- Exclude: 14 notes missing sex.
- Exclude: 15,770 notes where FSA not in the set $\{M,L\}$.
- Exclude: 340 notes if physician write $<$ 1000 notes (between 01/01/2017-31/12/2020).
- Exclude: 20,298 notes with $<$3 tokens or $>$ 150 tokens.

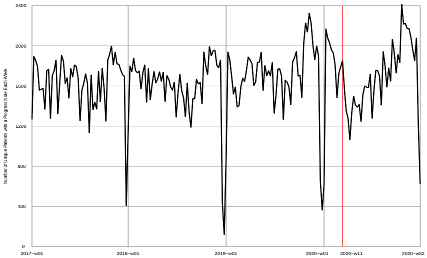
Sample Descriptive Statistics: Note-Level Unit of Analysis

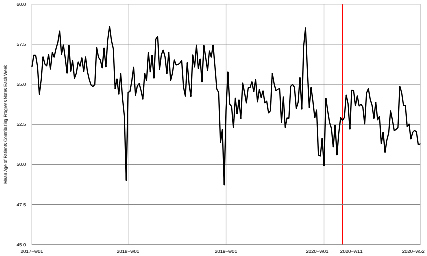
- 382,666 unique progress notes.
- 44,828 unique patients.
- 54 unique physicians.
- 12 distinct clinical practice sites across North-Central Toronto, Canada.

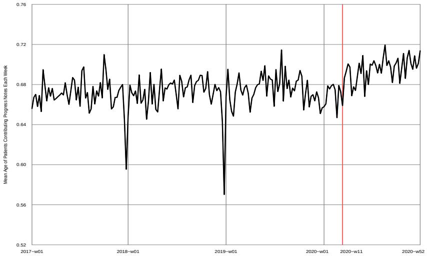
Variable	Variable Level	Count	Percentage
Age	0-20 years	36,344	10%
	20-40 years	71,481	19%
	40-65 years	130,172	34%
	65-85 years	112,293	29%
	85+ years	32,376	8%
Sex	Female	259,573	68%
	Male	123,093	32%
Year	2017	91,973	24%
	2018	91,906	24%
	2019	97,673	26%
	2020	101,11	26%

Number of Notes, Number of Patients, Mean Age, and Sex Over Time









Computationally Processing Digital Text Data

- No perfect pipeline to computationally process text data for clinical research.
- Computational methods must be fit for purpose, aligning with research objectives.

Overview of Method: simple, transparent, lightweight, scalable. Limitations?

- Tokenization on whitespace characters (\s, \t, \r, \f).
- Normalization (lowercase conversion, remove non-alphabetic characters).
- Review token dictionary, and manually curate a clinical vocabulary (P=2210).
- Bag of words, term-frequency vectors, and the document-term-matrix (DTM).
- Semantics, themes/topics, and the DTM representation/structure.

[DEMO: Tokenization, Normalization, and DTM creation]

Turney & Pantel. (2010). Frequency to Meaning: Vector Space Models of Semantics. JAIR.

Corpus Descriptive Statistics:

- 382,666 notes, 2210 unique tokens, 10,574,614 total tokens. DTM sparsity: 99.1%.
- Average note length (27.6 tokens) (Median=23; IQR=13-36).

Unigram Frequency Statistics for Overall and Manually Curated Corpus

Original Unprocessed Corpus			Manually Curated Corpus		
to (7.6M)	a (6.8M)	and (6.1M)	pain (316k)	bp (234k)	mg (225k)
for (5.5M)	no (4.9M)	by (4.4M)	back (134k)	work (104k)	feels (97k)
of (4.3M)	the (4.2M)	in (3.5M)	fever (87k)	chest (83k)	symptoms (79k)
on (3.1M)	with (2.7M)	off (2.5M)	meds (78k)	weight (72k)	blood (70k)
mg (2.0M)	is (1.9M)	or (1.9M)	systolic (67k)	heart (66k)	tablets (65k)
not (1.9M)	at (1.7M)	updated (1.6M)	diastolic (65k)	flu (64k)	bw (63k)
you (1.5M)	her (1.4M)	she (1.4M)	tablet (62k)	cough (60k)	feeling (59k)
has (1.4M)	if (1.3M)	as (1.3M)	sleep (58k)	meds (57k)	referral (55k)
this (1.2M)	will (1.2M)	that (1.2M)	bpm (52k)	sx (51k)	anxiety (50k)
po (1.2M)	tabs (1.2M)	nsigned (1.1M)	rx (49k)	mood (48k)	vaccine (47k)
was (1.1M)	have (1.0M)	normal (992k)	dose (45k)	tylenol (45k)	shot (44k)
pain (967k)	pt (964k)	ncreated (959k)	family (43k)	swelling (42k)	abdo (42k)
direct (940k)	refills (938k)	but (845k)	knee (41k)	skin (41k)	rn (40k)
be (836k)	office (819k)	tablet (817k)	throat (40k)	er (39k)	diet (39k)
daily (817k)	dr (797k)	he (765k)	covid (39k)	exercise (38k)	neck (38k)
are (761k)	had (733k)	patient (732k)	health (38k)	ear (38k)	urine (36k)
days (731k)	today (722k)	from (698k)	felt (36k)	pap (35k)	med (34k)

Section 3: DTMs, Matrix Factorization, and Topic Modelling

- Word-Frequency Statistics. Semantics and Meaning. The Document Term Matrix.
- Non-Negative Matrix Factorization Topic Models.
- Characterizing Important/Meaningful Primary Care Practice Patterns.

UNIVERSITY OF TORONTO

Matrix factorization/decomposition of input DTM

- Tokenize input clinical text. Represent text as "bag of words".
- Document Term Matrix: rows are document specific term-frequency vector.
- Document Term Matrix: N documents. P words in vocabulary.
- NMF: k=1...K rows of ϕ represent topical/archetypical vectors.
- NMF: n=1...N rows of θ represent patient affinity to topics.
- Complexity of model vs. quality of fit governed by hyper-parameter (K=1...min(N,P)).

$$\overset{N \times P \text{ matrix}}{\begin{bmatrix} x_{1,1} & \cdots & \cdots & \cdots & x_{1,P} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N,1} & \cdots & \cdots & \cdots & x_{N,P} \end{bmatrix}} \approx \overset{N \times K \text{ matrix}}{\begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,K} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \theta_{N,1} & \cdots & \theta_{N,K} \end{bmatrix}} * \overset{K \times P \text{ matrix}}{\begin{bmatrix} \phi_{1,1} & \cdots & \cdots & \cdots & \phi_{1,P} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{K,1} & \cdots & \cdots & \cdots & \phi_{K,P} \end{bmatrix}}$$

[DEMO: NMF and Low Rank Approximations]

Lee et al, (1999). NMF: Finding Parts in Objects. Nature.
Blei et al, (2003). Latent Dirichlet Allocation. JMLR.
Deerwester et al. (1990). Latent Semantic Analysis. JASIS.

Non-Negative Matrix Factorization/Decomposition of input DTM

- DTM is high-dimensional, sparse, non-negative, and over-determined.
- Many strong pair-wise correlations amongst column vectors (words).
- Many strong pair-wise correlations between row vectors (notes).
- Small number latent topical domains/bases needed to approximate DTM structure.
- Latent topical basis reveals practice patterns (roles, activities, patient-types).

Outer Product Representation of a Matrix Factorization Model

$$X \approx \sum_{k=1}^{K} \theta_{:k}\phi_{k:} = \begin{bmatrix} \theta_{11} \\ \vdots \\ \theta_{N1} \end{bmatrix} [\phi_{11} \ldots \phi_{1P}] + \begin{bmatrix} \theta_{12} \\ \vdots \\ \theta_{N2} \end{bmatrix} [\phi_{21} \ldots \phi_{2P}] + \cdots + \begin{bmatrix} \theta_{1K} \\ \vdots \\ \theta_{NK} \end{bmatrix} [\phi_{K1} \ldots \phi_{KP}]$$

[DEMO: Interactively Browsing Topical Practice Patterns at NYFHT]

UNIVERSITY OF TORONTO

Exploration of Latent Primary Care Topical/Thematic Vectors (1-25)

Topic	Token 1	Token 2	Token 3	Token 4	Token 5
Topic 1	tylenol (0.35)	advil (0.09)	tab (0.03)	headache (0.03)	tabs (0.02)
Topic 2	mg (0.45)	tab (0.02)	tabs (0.02)	capsules (0.01)	po (0.01)
Topic 3	fever (0.34)	diarrhea (0.03)	vomiting (0.02)	tylenoladvil (0.02)	viral (0.02)
Topic 4	neck (0.21)	head (0.04)	arm (0.03)	headache (0.02)	headaches (0.02)
Topic 5	bw (0.31)	iron (0.03)	tsh (0.02)	ferritin (0.02)	thyroid (0.02)
Topic 6	work (0.47)	social (0.04)	stress (0.03)	working (0.03)	treatment (0.03)
Topic 7	bp (0.58)	systolic (0.04)	diastolic (0.03)	htn (0.03)	norvasc (0.03)
Topic 8	sleep (0.37)	bed (0.05)	sleeping (0.03)	apnea (0.02)	insomnia (0.02)
Topic 9	anxiety (0.3)	anxious (0.04)	panic (0.03)	social (0.02)	counselling (0.02)
Topic 10	flu (0.37)	shot (0.32)	anaphylactic (0.03)	influenza (0.03)	ibuprofen (0.02)
Topic 11	weight (0.32)	kg (0.09)	bmi (0.05)	ht (0.04)	lbs (0.03)
Topic 12	pain (0.52)	palpation (0.02)	flexion (0.01)	physio (0.01)	arm (0.01)
Topic 13	ear (0.31)	hearing (0.06)	ears (0.05)	wax (0.05)	cerumen (0.05)
Topic 14	eating (0.05)	diet (0.04)	food (0.04)	wt (0.03)	snack (0.03)
Topic15	throat (0.23)	sore (0.13)	strep (0.04)	viral (0.03)	nodes (0.03)
Topic 16	rx (0.43)	shingrix (0.01)	ativan (0.01)	ra (0.01)	abx (0.01)
Topic 17	meds (0.43)	bmd (0.01)	vit (0.01)	chronic (0.01)	bone (0.01)
Topic 18	pap (0.12)	bleeding (0.04)	vaginal (0.03)	discharge (0.02)	pelvic (0.02)
Topic 19	vaccine (0.20)	influenza (0.08)	flu (0.08)	allergy (0.06)	fever (0.05)
Topic 20	dose (0.31)	medication (0.10)	immunization (0.05)	injection (0.04)	shingrix (0.03)
Topic 21	breast (0.27)	cancer (0.03)	nipple (0.03)	mammogram (0.02)	lump (0.02)
Topic 22	medications (0.15)	allergy (0.06)	drug (0.05)	capsules (0.05)	capsule (0.05)
Topic 23	cough (0.26)	sob (0.03)	ventolin (0.03)	asthma (0.03)	coughing (0.03)
Topic 24	bilat (0.26)	masses (0.02)	neuro (0.02)	limbs (0.02)	head (0.02)
Topic 25	heart (0.20)	bpm (0.17)	systolic (0.16)	diastolic (0.16)	bp (0.02)

UNIVERSITY OF TORONTO

Exploration of Latent Primary Care Topical/Thematic Vectors (26-50)

Topic	Token 1	Token 2	Token 3	Token 4	Token 5
Topic 26	urine (0.14)	uti (0.07)	urinary (0.04)	dysuria (0.04)	hematuria
Topic 27	eye (0.28)	vision (0.06)	drops (0.05)	eyes (0.04)	discharge (0.04)
Topic 28	symptoms (0.42)	nausea (0.02)	urinary (0.02)	headache (0.01)	gi (0.01)
Topic 29	foot (0.12)	swelling (0.07)	ankle (0.04)	toe (0.04)	feet (0.02)
Topic 30	sx (0.41)	neuro (0.03)	gi (0.03)	urinary (0.02)	melena (0.02)
Topic 31	mother (0.30)	father (0.05)	parents (0.02)	sister (0.02)	mothers (0.02)
Topic 32	mood (0.22)	cipralex (0.04)	depression (0.03)	counselling (0.03)	speech (0.03)
Topic 33	exercise (0.06)	diet (0.05)	ldl (0.03)	screening (0.02)	cancer (0.02)
Topic 34	tablets (0.27)	tablet (0.26)	medications (0.07)	oral (0.05)	mg (0.04)
Topic 35	rn (0.24)	immunization (0.03)	injection (0.03)	baby (0.02)	arm (0.02)
Topic 36	er (0.24)	felt (0.05)	head (0.03)	ct (0.03)	sob (0.02)
Topic 37	covid (0.23)	health (0.14)	physical (0.13)	emergency (0.11)	pandemic (0.04)
Topic 38	back (0.49)	spine (0.02)	lumbar (0.02)	flexion (0.02)	physio (0.02)
Topic 39	mom (0.36)	dad (0.03)	parents (0.02)	baby (0.02)	feeding (0.01)
Topic 40	chest (0.27)	sob (0.04)	cvs (0.03)	edema (0.02)	palpitations (0.02)
Topic 41	knee (0.29)	swelling (0.05)	oa (0.03)	joint (0.03)	medial (0.03)
Topic 42	blood (0.31)	pressure (0.14)	medication (0.03)	pulse (0.03)	pounds (0.02)
Topic 43	family (0.08)	social (0.06)	counselling (0.04)	husband (0.04)	daughter (0.04)
Topic 44	feeling (0.40)	felt (0.05)	tired (0.03)	anxious (0.03)	treatment (0.02)
Topic 45	feels (0.50)	felt (0.03)	tired (0.01)	stress (0.01)	anxious (0.01)
Topic 46	hip (0.23)	xray (0.05)	oa (0.03)	physio (0.03)	flexion (0.02)
Topic 47	nasal (0.19)	sinus (0.06)	congestion (0.06)	nose (0.04)	nasonex (0.03)
Topic 48	skin (0.13)	rash (0.08)	cream (0.04)	derm (0.03)	lesions (0.03)
Topic 49	referral (0.32)	derm (0.03)	ent (0.02)	gi (0.02)	mri (0.01)
Topic 50	abdo (0.13)	diarrhea (0.04)	stool (0.03)	bm (0.03)	masses (0.03)

Section 4: Temporal Topic Models and COVID-19 Monitoring/Evaluation

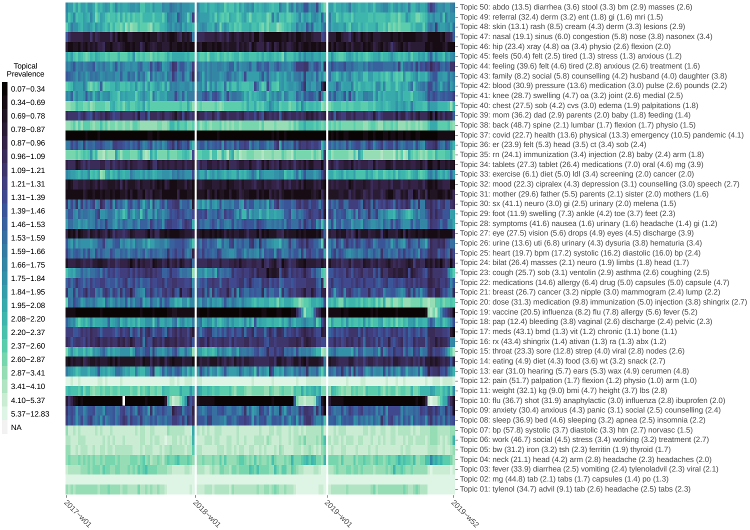
- A simple multivariate transformation for temporal topic modelling.
- Characterizing and monitoring the evolution of primary care topics over time.
- Identification and evaluation of COVID-19 pandemic effects on primary care topical series.

Temporal Topic Modelling

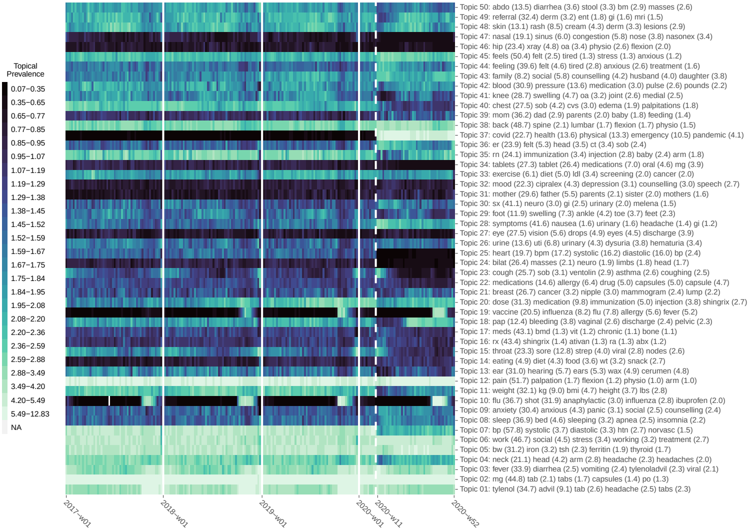
- Observe N*P dimensional DTM (X). Observe meta-data (z_i) for each note (i=1...N).
- Assume meta data is discrete (e.g. time, space, sex/age, physician ID, clinic ID, etc.).
- Assume, for each note i=1...N we have $z_i \in (1 \ldots T)$. T discrete levels.
- Realize, each note i=1...N is associated with length K topical prevalence vector.
- For each stratifying factor (t=1...T), compute length-K mean topical prevalence vector.
- Resulting T*K dimensional MVTS object used for evaluating COVID-19 pandemic effects.

$$
\begin{array}{c}
z_i(t=1)\left\{\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right. \\
z_i(t=2)\left\{\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right. \\
z_i(t=3)\left\{\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right. \\
z_i(t=T)\left\{\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right.
\end{array}
\begin{bmatrix}
x_{1,1} & \cdots & \cdots & x_{1,P} \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
x_{N,1} & \cdots & \cdots & x_{N,P}
\end{bmatrix}
\approx
\begin{bmatrix}
\theta_{1,1} & \cdots & \theta_{1,K} \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\theta_{N,1} & \cdots & \theta_{N,K}
\end{bmatrix}
*
\begin{bmatrix}
\phi_{1,1} & \cdots & \cdots & \phi_{1,P} \\
\vdots & \vdots & \vdots & \vdots \\
\phi_{K,1} & \cdots & \cdots & \phi_{K,P}
\end{bmatrix}
$$

Monitoring Evolution of Primary Care Topical Time Series (2017-2019)

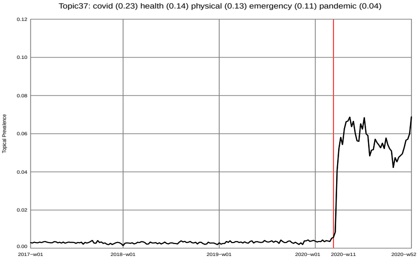


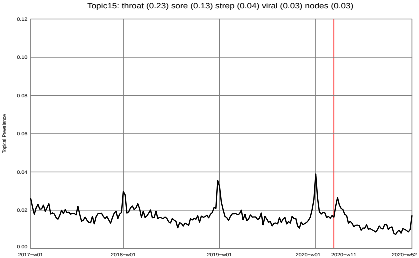
Identifying and Understanding COVID-19 Pandemic Effects (2017-2020)

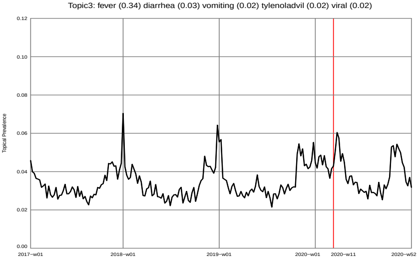


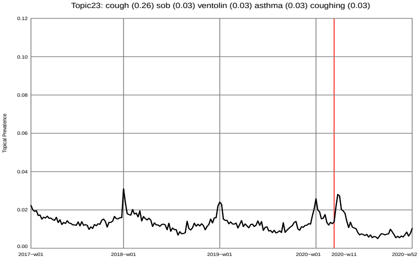
UNIVERSITY OF TORONTO

COVID-19 Impacts on Established Seasonal Harmonic Patterns

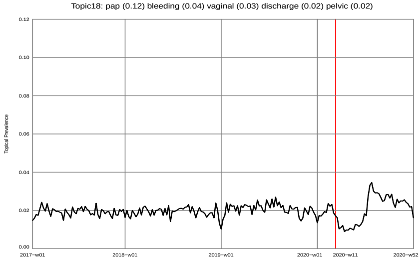


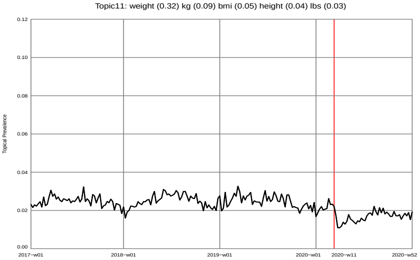


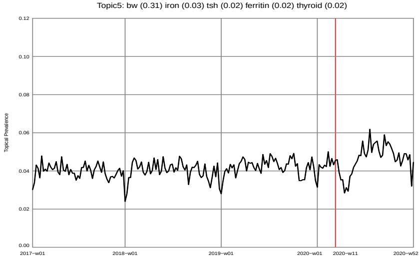


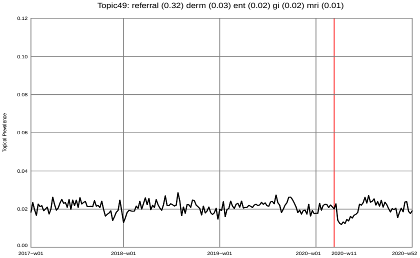


COVID-19 Induces Short-term Impacts and Compensatory Effects

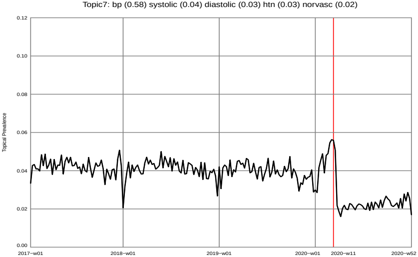


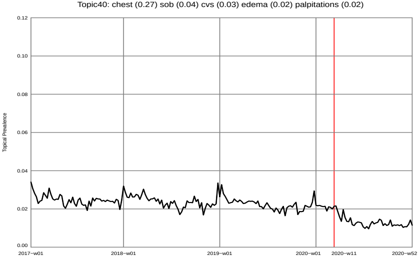


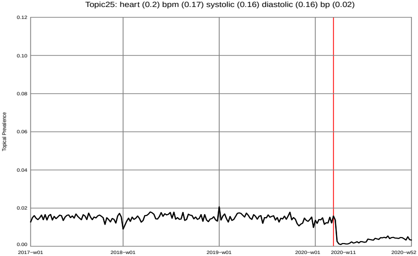


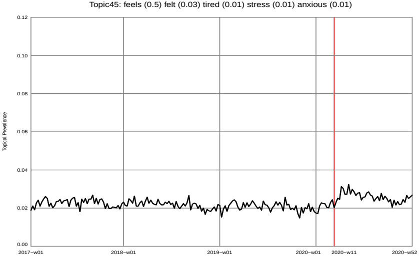


COVID-19 Results in Immediate and Sustained Impacts

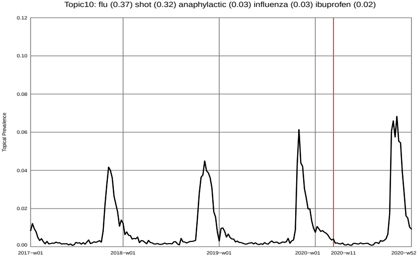


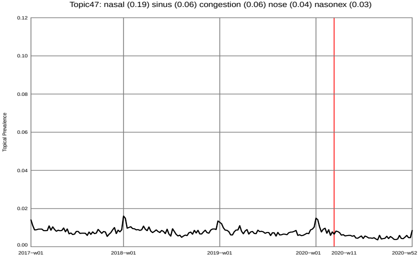


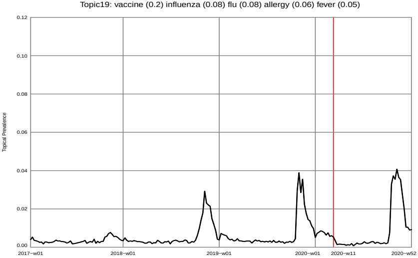


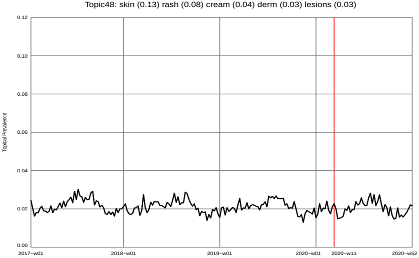


Topical Dynamics Unchanged over COVID-19 Pandemic









Conclusions

- Clinical and methodological conclusions.
- Limitations of proposed research and resulting inferences.
- Opportunities for future work.

Conclusions

- NMF learns meaningful topical basis characterizing primary care practice patterns.
- Temporal topic models identify COVID-19 impact on certain thematic domains:
    - Management of physical/mental health concerns during COVID-19 pandemic.
    - Symptoms of viral upper respiratory tract infection (e.g. fever, cough, sob, etc.).
    - CDPS, labs/vitals/meds, specialist referrals, family/work, anxiety, etc..

From Text to Topics to Health System and Community Health Insights

- We observe changes in latent topical series (especially following the COVID-19 pandemic).
- Changes in topics reflect changes in word-frequency utilization patterns/statistics.
- Changes in word-frequency statistics reflect evolving primary care practice patterns.

Methodological Thoughts...

- Primary care text data captures unique information not available in other data sources.
- Unsupervised monitoring of text data identifies COVID-19 impacts on community health.
- Design/methodology is cost-effective, scalable, yields unique/meaningful insights.
- More research needed on evaluating unsupervised designs for COVID-19 monitoring.

Limitations

- Primary care data. Not representative of LTC-, ED-, ICU-settings.
    - Only text data. Characterization incomplete if data recorded elsewhere in EMR.
- Informative visit bias. Patient selection bias. Physician recording biases.
- Sensitivity analyze alternative computational string processing pipelines:
    - NLTK, spacy, clinspacy, medspacy, cTAKES, NILE, etc.
- Subjective aspect of manually curated dictionary of clinical tokens.
    - Map onto validated nomenclature/ontology: SNOMED, UMLS, OMOP, etc.
    - Physician curation. Inter-rater agreement. Delphi approach to consensus dictionary.
- Model selection based on qualitative/descriptive evaluation.
    - Quantitative evaluation metric, CV-evaluation over grid model hyper-parameters.
    - What methods exist for "validating" quality of unsupervised phenotypes?
- Alternative model classes: dynamic-LDA, structural topic models, seq-NMF, etc.
- Current models ignore complex hierarchical, temporal, etc. dependencies.
- Descriptive target of inference.
    - No estimates uncertainty. No tests of intervention effects. Etc.
    - Conclusions are hypothesis generating and require further validation/triangulation.

Future Clinical/Methodological Work:

- Translating information to key stakeholders: who and how?
- Analyze UTOPIAN Q2-2021 data; evaluating COVID-19 impacts (Wave-02 and Wave-03).
- POPLAR text data? Evaluate COVID-19 impacts regionally across province of Ontario?
- CPCSSN/HDRN national/Canadian evaluation? Federated NMF algorithms?

Incorporating Text Data into Primary Care Research Studies

- How have others used primary care clinical text data in research studies?
- Alternative sources of primary care text data (e.g. education, research, etc.)?
- Generating/curating open primary care NLP resources/datasets?
- Collaborative opportunities (projects, grants, etc.) focused on primary care NLP?

Thank You

christopher.meaney@utoronto.ca