Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Non-negative matrix factorization temporal topic models and clinical text data identify COVID-19 pandemic effects on primary healthcare and community health in Toronto, Canada

Christopher Meaney [a,b,*], Michael Escobar [b], Rahim Moineddin [a,b,d], Therese A. Stukel [c,d], Sumeet Kalia [a,b], Babak Aliarzadeh [a], Tao Chen [a], Braden O'Neill [a], Michelle Greiver [a,e]

[a] Department of Family and Community Medicine, University of Toronto, Canada
[b] Dalla Lana School of Public Health, University of Toronto, Canada
[c] IHPME, University of Toronto, Canada
[d] ICES, Toronto, Canada
[e] Department of Family and Community Medicine, North York General Hospital and University of Toronto, Canada

ARTICLE INFO

Keywords:
COVID-19
Electronic medical record
Text mining
Temporal topic modelling
Non-negative matrix factorization
Time series analysis

ABSTRACT

Objective: To demonstrate how non-negative matrix factorization can be used to learn a temporal topic model over a large collection of primary care clinical notes, characterizing diverse COVID-19 pandemic effects on the physical/mental/social health of residents of Toronto, Canada.

Materials and Methods: The study employs a retrospective open cohort design, consisting of 382,666 primary care progress notes from 44,828 patients, 54 physicians, and 12 clinics collected 01/01/2017 through 31/12/2020. Non-negative matrix factorization uncovers a meaningful latent topical structure permeating the corpus of primary care notes. The learned latent topical basis is transformed into a multivariate time series data structure. Time series methods and plots showcase the evolution/dynamics of learned topics over the study period and allow the identification of COVID-19 pandemic effects. We perform several post-hoc checks of model robustness to increase trust that descriptive/unsupervised inferences are stable over hyper-parameter configurations and/or data perturbations.

Results: Temporal topic modelling uncovers a myriad of pandemic-related effects from the expressive clinical text data. In terms of direct effects on patient-health, topics encoding respiratory disease symptoms display altered dynamics during the pandemic year. Further, the pandemic was associated with a multitude of indirect patient-level effects on topical domains representing mental health, sleep, social and familial dynamics, measurement of vitals/labs, uptake of prevention/screening maneuvers, and referrals to medical specialists. Finally, topic models capture changes in primary care practice patterns resulting from the pandemic, including changes in EMR documentation strategies and the uptake of telemedicine.

Conclusion: Temporal topic modelling applied to a large corpus of rich primary care clinical text data, can identify a meaningful topical/thematic summarization which can provide policymakers and public health stakeholders a passive, cost-effective, technology for understanding holistic impacts of the COVID-19 pandemic on the primary healthcare system and community/public-health.

1. Introduction

The coronavirus disease 2019 (COVID-19) pandemic is associated with the viral respiratory pathogen severe acute respiratory syndrome coronavirus-2 (SARS-CoV2). COVID-19 infection is known to cause a spectrum of acute symptoms and long-term health consequences [1,2]. Some infected individuals experience asymptomatic or mild disease [3,4]. Many COVID-19 infections result in moderate disease, with patients displaying symptoms consistent with acute respiratory disease and/or viral pneumonia (fever, cough, shortness of breath, sore throat, headache, tiredness) [5]. A small proportion of COVID-19 infections result in severe, life-threatening symptoms (e.g. acute respiratory

[*] Corresponding author.
E-mail address: christopher.meaney@utoronto.ca (C. Meaney).

https://doi.org/10.1016/j.jbi.2022.104034
Received 17 September 2021; Received in revised form 14 January 2022; Accepted 18 February 2022
Available online 22 February 2022

distress syndrome) or death [6]. Patients can experience a spectrum of symptoms long after recovering from COVID-19 infection (post-acute COVID-19 syndrome and “long COVID”) [1].

Across the globe, governments in concert with their public health institutions have experimented with a multitude of non-pharmaceutical interventions to mitigate the spread of COVID-19 within their jurisdictions [7,8]. Large scale public health communication efforts have promoted the importance and benefits of social distancing, mask wearing and hand washing as simple effective measures to reduce COVID-19 transmission risks. Contact tracing has been used to identify and quarantine infected persons and their contacts. In certain circumstances, governments have implemented targeted “stay-at-home orders”, and other “lockdown measures”. Oftentimes, these lockdown orders include a combination of non-essential business closures, school closures, limits on large scale gatherings, closure of government and civic institutions, and restricted inter-jurisdictional travel. These blunt policy-instruments aim to directly reduce viral disease transmission through reduced person-to-person contact. The indirect consequences of these policy interventions can include side-effects, such as social isolation and disrupted familial relationships, reduced employment and loss of income, decreased access to essential services including healthcare, and changes to health behaviours, lifestyle factors (e.g. diet/food-security, exercise), substance use patterns, and overall physical and mental health [9]. COVID-19 can impact individual and population health in many ways; and an improved understanding of the myriad of mechanisms through which the COVID-19 pandemic has impacted lives will help in planning for a post-COVID recovery.

To date, primary care clinical text data has been a relatively underutilized data source for evaluating and monitoring COVID-19 pandemic effects. With increased adoption of electronic medical record (EMR) systems, primary care physicians are amassing large quantities of longitudinal digital health information on their patients. A large proportion of the patient EMR data exists in an unstructured free-text format captured in clinical notes. The clinical narratives consist of rich and expressive information capturing important patient information and activities performed by the primary healthcare system, providing unique insights regarding patient symptoms, disease diagnosis and prognosis, and disease prevention and management. This includes the management of COVID-19 in community-based primary care. We posit that primary care practice patterns will have evolved over the pandemic to address changing local and community needs and EMR clinical text data can be used to characterize this evolution.

The objective of this study is to demonstrate how unsupervised temporal topic models can be applied to a large body of primary care clinical text data, to generate a meaningful characterization of the impact of the COVID-19 pandemic on the primary healthcare system in Toronto, Canada. We illustrate how the learned topic model characterizes dominant activities performed by primary care physicians; and encodes acute and chronic health conditions commonly encountered in patients. Given the learned latent topical basis over the corpus, we apply a straightforward multivariate data transformation to the resulting parameter estimates, yielding a multivariate time series data structure. Using time series methods and visualizations, we monitor the evolution of primary care topical series over time, and we identify direct and indirect COVID-19 pandemic effects on primary care practice patterns. We conduct several robustness checks to illustrate the stability of the resulting inferences across model hyper-parameter configurations and/or data perturbations. We posit that the proposed methodology can be used as a cost-effective technology for passive surveillance of the primary health care system, characterizing changing primary healthcare practice patterns during the pandemic and identifying a myriad of direct and indirect pandemic effects on community physical, mental, and social health.

2. Methods

2.1. Study setting and context

The study setting was Toronto, Canada – the fourth most populous city in North America [10]. The study was conducted by the University of Toronto Practice Based Research Network (UTOPIAN, https://www.dfcm.utoronto.ca/landing-page/utopian) using free-text data from North York Family Health Team (NYFHT, https://www.nyfht.com). NYFHT is one of the largest primary care teams in Canada; 91 family physicians are members, with over 90,000 patients. These family physicians work in twelve clinic locations geographically distributed across the north-central part of Toronto, Canada. NYFHT provides primary healthcare services to patients living in some of Canada’s most densely populated, ethnically, and economically diverse and multicultural neighborhoods.

As of writing, Toronto, Canada has experienced three waves of COVID-19 infection since the WHO declared COVID-19 a global pandemic in March 2020. The first wave of COVID-19 infections occurred between March-2020 and July-2020. The second wave of COVID-19 infections occurred between September-2020 and March-2021. A third wave of COVID-19 began in March-2021 and has resolved in early June-2021 [11]. This study focuses on data extracted from the EMR up to December 31, 2020.

Toronto has been considered a “hot spot” for COVID-19 infections in Canada. It had high per-capita rates of lab-confirmed COVID-19 infections, and extensive government-imposed lockdown measures. Each pandemic wave has been associated with attempts to reduce person-to-person contacts and population mobility through “stay-at-home orders”, “shutdowns” or “lockdowns”, where non-essential services operated at reduced capacity or were closed entirely. Specific restrictions have included non-essential business closures, work/learn from home mandates, reduced access (or closures) of many public/civic services (e.g., schools, daycares, libraries, etc.) and limitations on inter-jurisdictional travel/mobility.

2.2. Study design & study population

The study employs a retrospective open cohort design. We included information from the records of physicians consenting to share their EMR data with the University of Toronto Practice Based Research Network (UTOPIAN) for the purposes of research and quality improvement. All patients contributing at least one primary care clinical note between Jan-01–2017 through Dec-31–2020 were included in the study sample. The clinical note was the primary unit of analysis. Clinical notes include digital narratives arising from in-person visits, emails, telephone calls and other communications, whether entered by the family physician or other member of the primary care team. Clinical notes were excluded if there was missing information on any of the following variables: physician ID, patient ID, note ID, note date, patient age, patient sex, or patient postal code. We only included clinical notes with forward sortation areas (the first elements of the postal code) beginning with M (Toronto, Canada) or L (Greater Toronto Area outside of the city of Toronto). We limit our analysis to primary care physicians who on average wrote at least one note per day over the study period.

2.3. Computationally processing the primary care clinical note text corpora

Raw text data is digitally stored as a flat file/list/vector (with n=1… N elements); each clinical note is a sequence of digital characters of varying length. We map each clinical note to a P-dimensional term-frequency vector. Mapping each clinical note (n=1…N) to a unique vector yields an N-by-P array of data $\left(X \in R^{NP}\right)$. A given element of the matrix, X_{np}, is a non-negative count variable, denoting the number of times a given token/word (p) occurs in each clinical note (n). The data structure is colloquially known as a “document-term-matrix” (DTM) [12,13].

The chief problem associated with construction of the DTM is specification of the set of tokens (p=1…P) included in the final dictionary of the corpus. Below we summarize key steps in our sequential text tokenization and normalization pipeline:

- We tokenize text strings on white space boundaries (i.e. spaces, tabs, newlines, etc.).
- We normalize tokens using lowercase conversion.
- We normalize tokens, keeping only alphabetic characters (removing digits/punctuation).
- We remove single character tokens from our final dictionary.
- We tabulate and sort remaining tokens by decreasing occurrence frequency, manually review the returned token list, and include only tokens corresponding to clinically relevant entities.

Following manual review we identified P=2210 distinct tokens for inclusion in the final vocabulary/dictionary. The total number of tokens in the reduced corpus was10,574,614. The tokens included were mainly medical terms (e.g. disease names, disease symptoms, drug names, medical procedures, medical specialties, anatomical locations, etc.). We excluded stop words/tokens (syntactic/functional tokens containing little clinical semantic meaning). Words with low occurrence frequency were excluded (to save time/resources in our manual review; and, for computational considerations related to NMF model fitting). The entire list of P=2210 tokens is included in supplemental Appendix A.

All text processing was conducted using 64-bit R version 4.0.2. We used the following base R string processing functions (i.e. no external packages) to carry out tokenization and vocabulary normalization: strsplit(), gsub(), nchar(), length(), table(), match().

2.4. Non-Negative matrix factorization and topic models

Non-Negative Matrix Factorization (NMF) is a statistical model which seeks to factorize or decompose a non-negative input matrix into two non-negative sub-matrices [14,15]. When NMF is applied to an input DTM, the resulting matrix factorization results in a topic model interpretation [16,17].

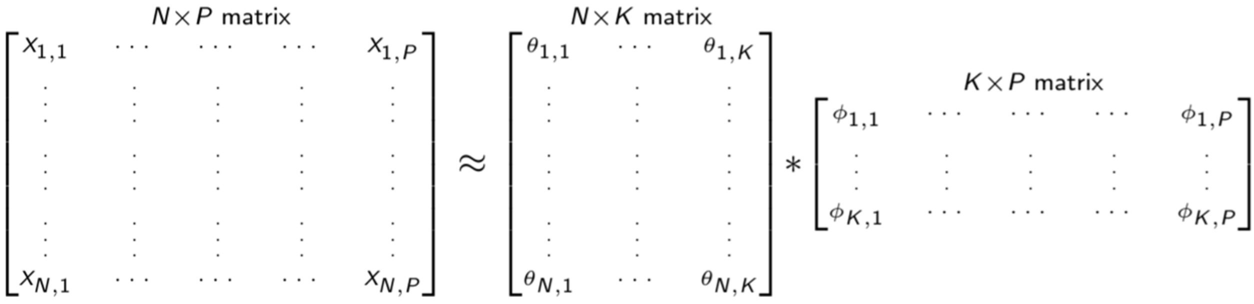
More specifically, NMF factorizes the N-by-P dimensional DTM into two sub-matrices of dimension N-by-K (θ) and K-by-P (ϕ), respectively. The DTM (X) consists of non-negative integers (i.e. word frequency counts) whereas the learned matrices ($\{\theta, \phi\}$) consist of non-negative real values. Mathematically the goal is to learn the values of the latent matrices (θ, ϕ) that best approximate the input dataset (X), subject to the non-negativity constraints ($X \approx \theta\phi$). Details on iterative estimation routines for NMF models are covered in [18,19].

The statistical quantities returned from fitting a NMF model to the input DTM include a N-by-K dimensional matrix of per-document topic weights/distributions (θ) and a K-by-P dimensional matrix of per-topic word weights/distributions (ϕ). The matrix of per-document topic weights/distributions (θ) act to indicate which topics are important to a given document. The matrix of per topic word weights (ϕ), describe a distribution over words/tokens empirically observed in the vocabulary. Individual vectors of the matrix (ϕ_k for k=1…K) cluster semantically correlated words/tokens and can be used to characterize the topical/thematic content of the document corpora.

Non-negative matrix factorization models were fit in 64-bit Python version 3.6 using the function sklearn.decomposition.NMF() from sklearn version = 0.24.2 module. Our primary analytic model contains K=50 latent topical/thematic bases. We do not regularize the latent parameter matrices. Parameter matrices are randomly initialized. Parameter matrices are updated using a gradient descent method on an Frobenius-norm loss function. We employ a loss function convergence tolerance of 1e-5 for terminating iterative updates.

2.5. Time series analysis and temporal topic modelling methodology

The matrix of per-note topical prevalence weights ($\theta \in R^{NK}$) provides a low-dimensional representation of each input clinical note in the corpus (n=1…N). After transformation, the NMF model can be thought of representing each clinical note as a K-dimensional categorical probability vector, expressing the affinity of a given note to a given latent topic. For each primary care clinical note in this study, we know the associated date between Jan-01–2017 through Dec-31–2020 which the note was recorded. We discretize the time scale into 208 weekly intervals (indexed t=1…T). In a given weekly interval, we observe a subsample N_t clinical notes, each represented in their K-dimensional topical space. For each interval t=1…T, we compute a multivariate (length-K) topical mean vector using the sample of N_t notes in the particular weekly interval. Following such a transformation, we arrive at a multivariate time series data structure ($\overline{\theta}_{TK}$). Each column (k=1…K) represents a length-T time series object, describing how the average topical prevalence varies over weekly intervals in our study. We plot each of the k=1…K topical time series functions, qualitatively exploring topical evolution/dynamics through time. We are interested in describing trends, seasonal harmonic patterns, and identifying/exploring COVID-19 pandemic effects in each of these time series. Additionally, we partition each latent topical series into two distinct strata: 1) 156 weekly measures observed between 01-Jan-2017 through 31-Dec-2019 (a pre-pandemic period); and 2) 52 weekly measures observed between 01-Jan-2020 through 31-Dec-2020 (the first COVID-19 pandemic year). Using data from 2017 to 2019 we fit an additive linear model including year/week as discrete covariates to each of the k=1…50 univariate time series (we use an AR-1 correlation structure to account for residual



dependence in the time series). For each of the k=1…50 series, we estimate/plot the mean topical prevalence and associated 95% confidence interval for each of the 156 weekly intervals in the pre-pandemic period. Using these models, we project/forecast topical prevalence estimates into the 2020 pandemic year. For each of the 52 weeks under consideration we compare observed versus predicted values and assess whether observed values are contained within model-implied 95% prediction limits. These analyses complement the descriptive time series plots, and assist stakeholders identify which univariate topical prevalence series are potentially "out-of-control" during the pandemic period. The results/visualizations are reported in appendices E1-E50.

2.6. Evaluating model stability and robustness

The primary analytic model reported in this study included K=50 latent topical bases. We arrive at the final model (of complexity K=50) using a human-in-the-loop approach, where subject matter experts and analysts review aspects of learned matrices (θ and ϕ); ultimately selecting a model which appropriately trades off complexity and quality (noting by Occam's razor, that we prefer simpler models; provided they fit the underlying data structures with reasonable fidelity). The subjective "eyeballing" approach is commonly employed in applied topic modelling communities, however, lacks a sense of empirical rigor. To enhance confidence in the robustness of proposed inferences, we fit several additional temporal topic models under different hyper-parameter configurations and/or data perturbations, illustrating stability of conclusions across a variety of model fits.

The most important hyper-parameter in any topic model is the choice of model complexity (K). If K is chosen to be too small, the model will lack capacity to provide a holistic summary of complex document collections; and returned topical vectors may combine semantically unrelated words/tokens. Conversely, if K is chosen to be too large, the returned topical vectors may be redundant and a parsimonious explanation of a complex phenomena may not be achieved. In this study, we report on a model with K=50 latent topical bases; however, in supplemental appendices B1-B6 we also report on models trained with different complexity parameters K={25,40,45,55,60,75}, assessing inferential stability across fitted models.

Latent parameter matrices in temporal topic models are learned using iterative algorithms (in this study we use gradient descent). Initialization conditions for latent parameter matrices often play an important role in determining final model estimates (as the model loss landscape is high-dimensional and non-convex). In supplemental appendices C1-C5 we fit models using five different random seeds (resulting in five different random initializations of θ and ϕ) and investigate stability/robustness of resulting multivariate time series as a function of the matrix initialization hyper-parameter.

Lastly, we bootstrap five separate DTMs (sampling documents with replacement, from the empirically observed corpus). Large-sample bootstrap theory suggests each bootstrap replicate sample will contain 63.2% of the unique documents contained in the original collection (whereas; the remaining 36.8% of documents will not be included for analysis in a given bootstrap replicate sample). For each of the five bootstrap replicate samples, we fit five separate NMF models (with five different initialization seeds). We construct five separate multivariate time series data structures (using the learned latent matrices, from the five independent NMF fits) and assess stability/robustness of inferences under bootstrap data perturbation. The results of the bootstrap stability analysis are presented in appendices D1-D5.

2.7. Research ethics

This study received ethics approval from North York General Hospital Research Ethics Board (REB ID: NYGH #20–0014).

Table 1
Descriptive statistics for primary care clinical note corpus (note level unit of analysis) and contributing patient sample (patient level unit of analysis).

	N = 382,666 Unique Notes	N = 44,828 Unique Patients
Age		
- 0 to 20 years	36,344 (9.5%)	6,412 (14.3%)
- 20 to 40 years	71,481 (18.7%)	10,704 (23.9%)
- 40 to 65 years	130,172 (34.0%)	16,117 (35.9%)
- 65 to 85 years	112,293 (29.3%)	9,411 (21.0%)
- >85 years	32,376 (8.5%)	2,184 (4.9%)
Sex		
- Male	123,093 (32.2%)	16,929 (37.8%)
- Female	259,573 (67.8%)	27,899 (62.2%)
Year*		
- 2017	91,973 (24.0%)	28,599
- 2018	91,906 (24.0%)	28,276
- 2019	97,673 (25.6%)	30,321
- 2020	101,114 (26.4%)	27,858

* Note: The number of unique patients in each year of study does not sum to N = 44,828 (number of unique patients in overall sample).

3. Results

3.1. Descriptive characteristics and corpus summary statistics

Our transformed and processed primary care clinical note corpus was comprised of 382,666 unique clinical notes, observed on 44,828 unique patients. The youngest patient in the sample was a few days old, and the oldest patient in the sample was 110 years old. Female patients are observed more frequently than male patients. Table 1 describes the characteristics of the corpus and our patient sample.

3.2. An NMF topic model of the primary care system in Toronto, Canada

The input DTM is of dimension 382,666 (rows/notes) by 2210 (columns/tokens). The corpus is comprised of 10,574,614 total tokens. The DTM is 99.1% sparse (i.e. contains few non-zero elements). The top-50 most frequently occurring words in the initial and final corpora are given in Table 2 below. The most frequently occurring words in the initial corpus are dominated by syntactic/functional words whereas, the most frequently occurring words in the manually curated dictionary include medical words with relatively precise semantic meanings (apart from issues of polysemy). We note the term "covid" appears as one of the top-50 most frequent words, in spite of the pandemic only commencing in 2020.

We fit a NMF model with 50 latent topical bases to the primary care clinical note corpus. Table 3 lists the top-10 words loading most strongly on each of the K=50 learned latent topics. Tokens loading strongly on a specific topic tend to be semantically correlated. We use the learned topical basis vectors to characterize the major activities performed by the primary care physicians, and further provide an unsupervised representation of community health/well-being of those residing in Toronto, Canada.

NMF learns a topical summarization that meaningfully characterizes a diverse array of clinical activities which the archetypical primary care physician orchestrates on a daily basis, including: disease prevention (e.g. lifestyle monitoring, cancer screening, annual influenza programs), monitoring and management of chronic physical and mental health conditions (e.g. depression/anxiety, cardiovascular disease, arthritis, etc.), management of acute physical health conditions (e.g. cough/cold/flu, urinary tract infections, lesions/swelling, localized pain, etc.), medication prescribing, medical specialist referral coordination and handling of other social/familial-dynamic issues (e.g. patients/families, work, school, life-events, etc.).

Table 2
Top-50 most frequently occurring tokens/words in the final analytic primary care clinical note corpora.

Final Processed Clinical Note Corpus		
N = 10,574,614 total tokens		
N = 2210 unique tokens		
Token	Occurrence Frequency	Percentage of Total Tokens
pain	316,634	2.99
bp	234,785	2.22
mg	225,084	2.13
back	134,809	1.27
work	104,672	0.99
feels	96,593	0.91
fever	87,352	0.83
chest	83,473	0.79
symptoms	78,641	0.74
medications	77,904	0.74
weight	72,072	0.68
blood	69,853	0.66
systolic	66,768	0.63
heart	65,568	0.62
tablets	65,231	0.62
diastolic	64,739	0.61
flu	63,821	0.60
bw	63,391	0.60
tablet	61,908	0.59
cough	59,615	0.56
feeling	58,823	0.56
sleep	58,224	0.55
meds	57,090	0.54
referral	54,869	0.52
bpm	51,979	0.49
sx	50,555	0.48
anxiety	50,398	0.48
rx	48,583	0.46
mood	48,029	0.45
vaccine	47,387	0.45
dose	44,956	0.43
tylenol	44,586	0.42
shot	44,329	0.42
family	43,117	0.41
swelling	41,702	0.39
abdo	41,582	0.39
knee	41,376	0.39
skin	41,041	0.39
rn	40,282	0.38
throat	40,177	0.38
er	39,119	0.37
diet	38,814	0.37
covid	38,502	0.36
exercise	38,465	0.36
neck	38,082	0.36
health	38,078	0.36
ear	37,501	0.35
urine	36,143	0.34
felt	35,531	0.34

3.3. *Time series analysis characterizing evolution of mean topical prevalence vectors in Toronto, Canada*

We transform the latent matrix of per-note topical prevalence vectors ($\theta \in R^{NK}$) into a multivariate time series data structure ($\overrightarrow{\theta}^{'} \in R^{KT}$). We use a heatmap to visualize the evolution of latent primary care topics over the study timeframe (Fig. 1). We note that year-over-year dynamics of the latent primary care topical series appear roughly stable between 2017 and 2019. Following March 2020, when COVID-19 “lockdown measures” were initiated, we observe a myriad of changes in select latent primary care topical series. Each of the K=50 independent time series plots, depicting the evolution of topical prevalence over time, is included in Appendix Figures E1-E50. Observed values from each latent topical series can be compared against AR1 dynamic regression model predicted values; and provide a sense of which series exhibit meaningful COVID-19 related impacts. In general, four qualitative patterns of change are observed following March-2020 COVID-19 restrictions in the K=50 latent topical series: 1) in certain topical series the COVID-19 pandemic has resulted in altered seasonal/harmonic patterns , 2) other series demonstrate short-term COVID-19 pandemic effects which temper over-time eventually returning to pre-COVID-19 baseline, 3) select series exhibit COVID-19 pandemic related level-shifts in topical prevalence which have not trended towards baseline levels, and finally 4) few series suggest the COVID-19 pandemic has resulted in little/no change.

3.4. *Robustness and stability analyses of learned NMF temporal topic models*

We perform an extensive series of post-hoc robustness analyses to demonstrate the stability of temporal topic model inferences generated in this study. All the robustness/stability analyses conducted follow a similar template: we alter important model hyper-parameters and/or stochastically perturb the underlying DTM, we re-estimate NMF models and their latent parameter matrices (θ and ϕ), and critically re-interpret the resulting multivariate time series data structures ($\overrightarrow{\theta}^{'} \in R^{KT}$) used for inferring COVID-19 effects on primary care and community health. Each re-analysis results in generation of a unique heatmap used for visualizing the multivariate topical time series data structure (containing K={25,40,45,50,55,60,75} rows/topics depending on the NMF model fit; and all containing T=208 weekly time strata). The exhibits generated in the sensitivity analyses replicate Fig. 1 illustrated above. Appendices B1-B6 display the results of varying the NMF model complexity parameter over the grid K={25,40,45,55,60,75}. Appendices C1-C5 display the results of varying random initialization seeds for the latent parameter matrices estimated by NMF. And appendices D1-D5 display the results of randomly perturbing the input DTM using a stochastic bootstrap methodology. Nearly all descriptive inferences derived from Fig. 1 above are corroborated over the 16 sensitivity analyses conducted in this section, building trust that the inferences generated by the study are stable/robust (i.e. the temporal dynamics of the latent topical series observed in the sensitivity analysis appear qualitatively similar to what was observed on the original analysis presented in Fig. 1).

4. Discussion

This study has demonstrated how NMF temporal topic modelling can be applied to a large corpus of primary care clinical notes to: 1) extract an informative latent thematic basis which meaningfully summarizes clinical practice patterns and identifies physical/mental/social-health issues permeating the large text dataset, and 2) characterize the evolution of the latent topics over time, hence, permitting a holistic understanding of how the COVID-19 pandemic impacted primary care practice patterns in Toronto, Canada. The rich primary care clinical note data capture a myriad of pandemic-related effects. In terms of direct effects, topics encoding respiratory disease symptoms display altered dynamics during the COVID-19 pandemic year. Study findings suggest that the COVID-19 pandemic was associated with indirect effects on topical domains representing mental health (mood and anxiety related conditions), sleep, social/familial dynamics, measurement of vitals/labs, uptake of prevention/screening maneuvers, and referrals to medical specialists. The primary healthcare system has simultaneously been fighting COVID-19 in its local community, all the while capturing rich and expressive clinical text data describing salient aspects of patient-physician clinical interactions. This study illustrates how temporal topic modelling can be used to extract a meaningful summarization from these large bodies of unstructured clinical text data; providing a complementary lens for understanding both direct/indirect pandemic related effects which can readily be scaled into a cost-effective technology for passive surveillance of the primary healthcare system and community/public-health (before/during/after the pandemic).

The NMF temporal topic modelling methodology used in this study

Table 3
Latent topical basis vectors learned from fitting a NMF model to our primary care clinical note corpus. Each independent topical basis vector (k=1…50) is normalized to represent a discrete probability distribution over words in the vocabulary (p=2210). The value in parentheses () next to a given word denotes the extent to which that word loads onto the specific topic.

Topic1: tylenol (0.35) advil (0.09) tab (0.03) headache (0.03) tabs (0.02) injection (0.02) barre (0.02) formaldehyde (0.02) headaches (0.02) arthritis (0.02)
Topic2: mg (0.45) tab (0.02) tabs (0.02) capsules (0.01) po (0.01) capsule (0.01) cipralex (0.01) norvasc (0.01) crestor (0.01) coversyl (0.01)
Topic3: fever (0.34) diarrhea (0.03) vomiting (0.02) tylenoladvil (0.02) viral (0.02) rash (0.02) fluids (0.02) immunization (0.02) drinking (0.02) sorenessredness (0.01)
Topic4: neck (0.21) head (0.04) arm (0.03) headache (0.02) headaches (0.02) massage (0.01) cervical (0.01) physio (0.01) hand (0.01) numbness (0.01)
Topic5: bw (0.31) iron (0.03) tsh (0.02) ferritin (0.02) thyroid (0.02) anemia (0.02) fatigue (0.02) liver (0.01) dm (0.01) synthroid (0.01)
Topic6: work (0.47) social (0.04) stress (0.03) working (0.03) treatment (0.03) msw (0.01) counselling (0.01) relationship (0.01) felt (0.01) boss (0.01)
Topic7: bp (0.58) systolic (0.04) diastolic (0.03) htn (0.03) norvasc (0.02) coversyl (0.02) sob (0.01) tru (0.01) dizziness (0.01) amlodipine (0.01)
Topic8: sleep (0.37) bed (0.05) sleeping (0.03) apnea (0.02) insomnia (0.02) hygiene (0.02) tired (0.02) zopiclone (0.02) melatonin (0.02) fatigue (0.01)
Topic9: anxiety (0.3) anxious (0.04) panic (0.03) social (0.02) counselling (0.02) gad (0.02) cipralex (0.02) depression (0.02) ativan (0.01) medication (0.01)
Topic10: flu (0.37) shot (0.32) anaphylactic (0.03) influenza (0.03) ibuprofen (0.02) acetaminophen (0.02) family (0.02) reaction (0.02) sorenessredness (0.02) barre (0.01)
Topic11: weight (0.32) kg (0.09) bmi (0.05) height (0.04) lbs (0.03) lb (0.02) feeding (0.02) head (0.02) appetite (0.01) baby (0.01)
Topic12: pain (0.52) palpation (0.02) flexion (0.01) physio (0.01) arm (0.01) chronic (0.01) wrist (0.01) joint (0.01) sitting (0.01) muscle (0.01)
Topic13: ear (0.31) hearing (0.06) ears (0.05) wax (0.05) cerumen (0.05) oil (0.03) drops (0.02) syringed (0.02) otitis (0.02) infection (0.02)
Topic14: eating (0.05) diet (0.04) food (0.04) wt (0.03) snack (0.03) dinner (0.02) foods (0.02) meals (0.02) meal (0.02) nutrition (0.02)
Topic15: throat (0.23) sore (0.13) strep (0.04) viral (0.03) nodes (0.03) swab (0.02) tonsils (0.02) cervical (0.02) nose (0.01) fluids (0.01)
Topic16: rx (0.43) shingrix (0.01) ativan (0.01) ra (0.01) abx (0.01) tabs (0.01) bc (0.01) medications (0.01) pills (0.01) cream (0.01)
Topic17: meds (0.43) bmd (0.01) vit (0.01) chronic (0.01) bone (0.01) renal (0.01) hypertension (0.01) bc (0.01) mass (0.01) htn (0.01)
Topic18: pap (0.12) bleeding (0.04) vaginal (0.03) discharge (0.02) pelvic (0.02) cervix (0.02) screening (0.02) iud (0.02) exam (0.02) vag (0.02)
Topic19: vaccine (0.2) influenza (0.08) flu (0.08) allergy (0.06) fever (0.05) acetaminophen (0.05) reaction (0.05) injection (0.05) vaccination (0.05) injectable (0.04)
Topic20: dose (0.31) medication (0.1) immunization (0.05) injection (0.04) shingrix (0.03) tylenoladvil (0.02) deltoid (0.02) tsh (0.01) synthroid (0.01) anaphylaxis (0.01)
Topic21: breast (0.27) cancer (0.03) nipple (0.03) mammogram (0.02) lump (0.02) ca (0.02) exam (0.02) breasts (0.02) cyst (0.02) mammo (0.02)
Topic22: medications (0.15) allergy (0.06) drug (0.05) capsules (0.05) capsule (0.05) allergies (0.04) mcg (0.04) oral (0.04) nondrug (0.02) ventolin (0.02)
Topic23: cough (0.26) sob (0.03) ventolin (0.03) asthma (0.03) coughing (0.03) viral (0.02) wheeze (0.02) crackles (0.02) uri (0.02) flovent (0.02)
Topic24: bilat (0.26) masses (0.02) neuro (0.02) limbs (0.02) head (0.02) murmur (0.02) nodes (0.01) bs (0.01) sob (0.01) numbness (0.01)
Topic25: heart (0.2) bpm (0.17) systolic (0.16) diastolic (0.16) bp (0.02) kg (0.02) edema (0.01) cvs (0.01) height (0.01) gaeb (0.01)
Topic26: urine (0.14) uti (0.07) urinary (0.04) dysuria (0.04) hematuria (0.03) infection (0.03) vaginal (0.03) kidney (0.03) discharge (0.02) macrobid (0.02)
Topic27: eye (0.28) vision (0.06) drops (0.05) eyes (0.04) discharge (0.04) conjunctivitis (0.02) swelling (0.02) eyelid (0.02) exam (0.02) optometrist (0.01)
Topic28: symptoms (0.42) nausea (0.02) urinary (0.02) headache (0.01) gi (0.01) vomiting (0.01) dizziness (0.01) urti (0.01) concussion (0.01) bilaterally (0.01)
Topic29: foot (0.12) swelling (0.07) ankle (0.04) toe (0.04) feet (0.02) plantar (0.02) nail (0.02) lateral (0.02) xray (0.02) erythema (0.01)
Topic30: sx (0.41) neuro (0.03) gi (0.03) urinary (0.02) melena (0.02) sob (0.01) uri (0.01) dm (0.01) wt (0.01) bms (0.01)
Topic31: mother (0.3) father (0.05) parents (0.02) sister (0.02) mothers (0.02) baby (0.02) brother (0.01) relationship (0.01) family (0.01) cancer (0.01)
Topic32: mood (0.22) cipralex (0.04) depression (0.03) counselling (0.03) speech (0.03) appetite (0.02) energy (0.02) phq (0.02) gad (0.02) depressed (0.01)
Topic33: exercise (0.06) diet (0.05) ldl (0.03) screening (0.02) cancer (0.02) diabetes (0.02) dm (0.02) crestor (0.01) vit (0.01) bmd (0.01)
Topic34: tablets (0.27) tablet (0.26) medications (0.07) oral (0.05) mg (0.04) bedtime (0.03) mcg (0.02) synthroid (0.01) tab (0.01) crestor (0.01)
Topic35: rn (0.24) immunization (0.03) injection (0.03) baby (0.02) arm (0.02) head (0.02) sleeping (0.02) wbv (0.02) feeding (0.01) screen (0.01)
Topic36: er (0.24) felt (0.05) head (0.03) ct (0.03) sob (0.02) headache (0.02) nausea (0.02) vomiting (0.02) bleeding (0.01) dizziness (0.01)
Topic37: covid (0.23) health (0.14) physical (0.13) emergency (0.11) pandemic (0.04) exam (0.04) outbreak (0.02) mental (0.01) ed (0.01) working (0.01)
Topic38: back (0.49) spine (0.02) lumbar (0.02) flexion (0.02) physio (0.02) legs (0.01) exercises (0.01) sitting (0.01) massage (0.01) muscles (0.01)
Topic39: mom (0.36) dad (0.03) parents (0.02) baby (0.02) feeding (0.01) milk (0.01) friends (0.01) sister (0.01) brother (0.01) head (0.01)
Topic40: chest (0.27) sob (0.04) cvs (0.03) edema (0.02) palpitations (0.02) stress (0.02) ecg (0.01) cardiac (0.01) breath (0.01) murmurs (0.01)
Topic41: knee (0.29) swelling (0.05) oa (0.03) joint (0.03) medial (0.03) xray (0.02) knees (0.02) injury (0.02) physio (0.02) effusion (0.02)
Topic42: blood (0.31) pressure (0.14) medication (0.03) pulse (0.03) pounds (0.02) stool (0.02) bleeding (0.02) hypertension (0.01) bm (0.01) sugar (0.01)
Topic43: family (0.08) social (0.06) counselling (0.04) husband (0.04) daughter (0.04) son (0.03) treatment (0.02) children (0.02) alcohol (0.02) relationship (0.02)
Topic44: feeling (0.4) felt (0.05) tired (0.03) anxious (0.03) treatment (0.02) nausea (0.01) eating (0.01) energy (0.01) dizzy (0.01) sleeping (0.01)
Topic45: feels (0.5) felt (0.03) tired (0.01) stress (0.01) anxious (0.01) husband (0.01) worried (0.01) friends (0.01) kids (0.01) working (0.01)
Topic46: hip (0.23) xray (0.05) oa (0.03) physio (0.03) flexion (0.02) spine (0.02) fracture (0.02) groin (0.02) bmd (0.01) surgery (0.01)
Topic47: nasal (0.19) sinus (0.06) congestion (0.06) nose (0.04) nasonex (0.03) mcg (0.03) sinusitis (0.02) saline (0.02) discharge (0.02) sinuses (0.02)
Topic48: skin (0.13) rash (0.08) cream (0.04) derm (0.03) lesions (0.03) itchy (0.03) eczema (0.02) dermatitis (0.01) erythema (0.01) betaderm (0.01)
Topic49: referral (0.32) derm (0.03) ent (0.02) gi (0.02) mri (0.01) chronic (0.01) surgery (0.01) colonoscopy (0.01) gyne (0.01) ct (0.01)
Topic50: abdo (0.13) diarrhea (0.04) stool (0.03) bm (0.03) masses (0.03) constipation (0.02) bms (0.02) vomiting (0.02) nausea (0.02) gi (0.02)

identifies a myriad of primary care domains which have been impacted by the COVID-19 pandemic. We divide the K=50 latent primary care topical series into 4 groups displaying similar temporal dynamics: 1) series which the COVID-19 pandemic has resulted in altered seasonal/harmonic patterns, 2) series demonstrating short-term COVID-19 pandemic effects which temper over-time eventually returning to baseline, 3) series exhibiting COVID-19 pandemic related level-shifts in topical prevalence which have not trended towards baseline levels, and finally 4) series for which the COVID-19 pandemic has resulted in little/no change.

Topic 37 (covid, health, physical, emergency, pandemic) demonstrates the most noticeable change in temporal dynamics pre-vs-post COVID-19. We note the string “covid” occurred over 38,000 times in the corpus. The observed “covid” topic likely encodes direct COVID-19 pandemic effects on primary care clinical operating procedures (e.g. implementation of COVID-19 screens prior to in-person visits) and/or changes in EMR recording practices (e.g. templating of COVID-19 related physical/mental health examinations). We also observed seasonal harmonic disruptions in primary care topical series characterizing symptoms of acute respiratory disease and viral infection, characteristic symptoms of COVID-19 infection (Topic 3: fever, diarrhea, vomiting; Topic 15 sore, throat, viral; Topic 23 cough, sob). The token “covid” is the most-probable word under Topic 37, and appears in the top-5% of most probable tokens for each of Topic 3, Topic 15, and Topic 23. These topical changes suggest a primary care role in managing or coordinating care for those with COVID-19. Alternatively, these may be reflecting changing levels of vigilance in primary care physicians with respect to screening for COVID-19 like disease. Other primary care topical series whose dynamics appear altered during the COVID-19 pandemic include series 28 (symptoms, nausea, gi, urinary), series 36 (er, felt, sob) and series 1 (allergies and headaches).

Short-term COVID-19 pandemic effects are observed in many primary care topical series. These effects seem to arise following initiation of lockdown measures in the city of Toronto (March 2020), with effects mediating in the summer of 2020 as social activities were permitted to return to normal. The model identifies subtle changes in topics related to mental health, sleep, stress, work and social/familial dynamics (Topic 6, Topic 8, Topic 9, Topic 39, Topic 43, Topic 44, and Topic 45). We note

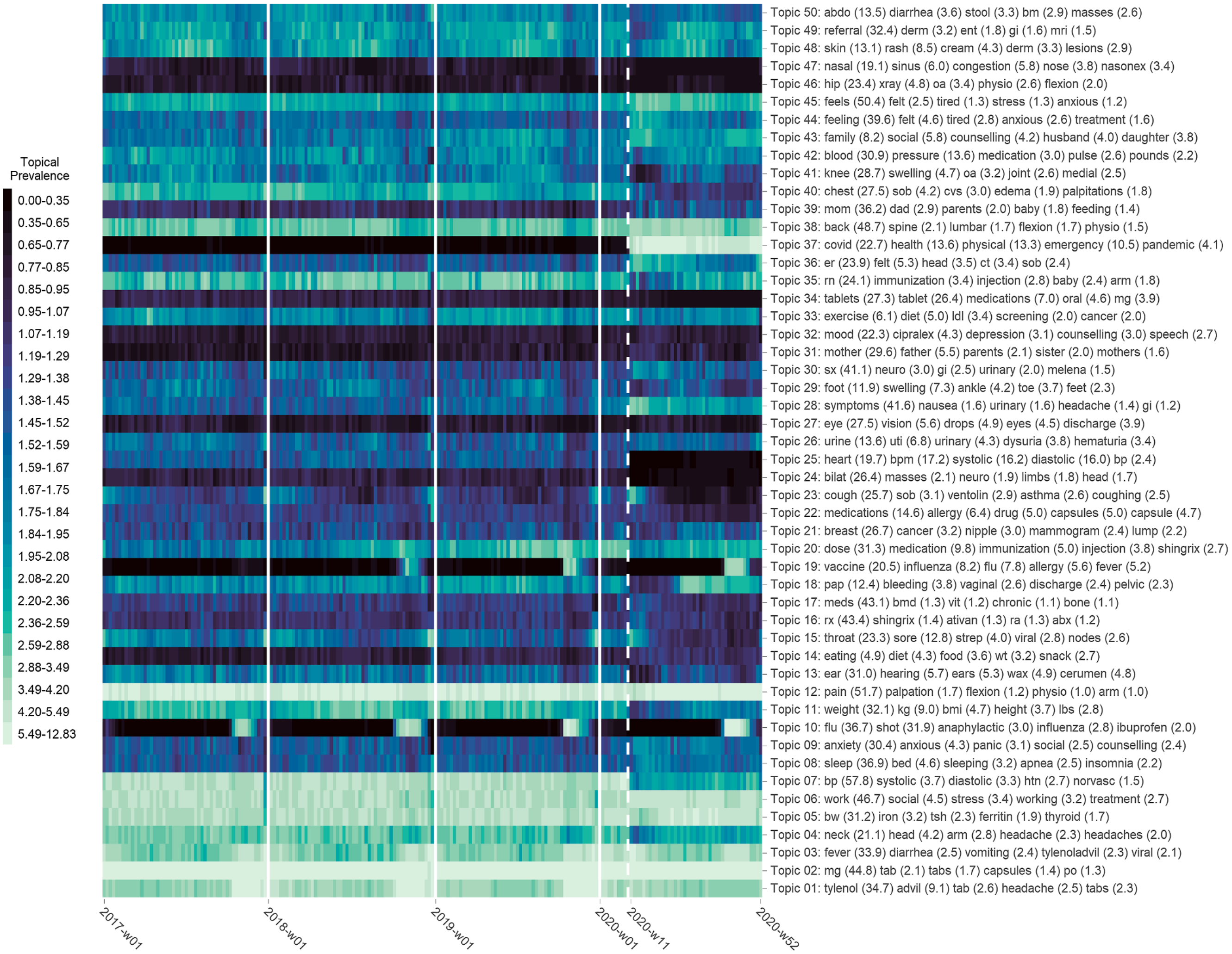


Fig. 1. A heatmap of the multivariate topical time series data structure (K=50 rows/topics; and T=208 weekly time strata). Each row of the heatmap denotes a distinct latent primary care topical time series. Color is used to signify variation in topical prevalence/importance over time. Solid vertical lines denote the start of a new calendar year (2017, 2018, 2019, 2020); and the dashed vertical line indicates the date the WHO declared COVID-19 a global pandemic (i.e. March 11, 2020).

that few other data sources, apart from primary care clinical text data, would contain rich/expressive information on mental health conditions and social/familial dynamics (or if the data were available from self-reported surveys, it would be subject to self-report bias, and lack the scientific/clinical precision of an expert primary care physician observer/monitor). Further, we notice short-term pandemic effects, followed by regression to baseline or over-compensation, in preventative care and screening topics (Topic 18: pap, bleeding, vaginal; Topic 20: injections/immunizations, cancer screening; and Topic 35: rn, immunization, baby). Similarly topics encoding labs/blood-work (Topic 5) and topics loading on referrals to medical specialists (Topic 45) also show short-term effects in line with wave-01 lockdowns; which reverse as restrictions are lifted. Finally, a number of topics characterizing chronic physical health conditions also illustrate short-term pandemic effects which remediate following the reversal of wave-01 lockdown measures (Topic 4: neuro, headache; Topic 12: pain, physio; Topic 13: ears, hearing; Topic 38: back, spine, physio; and Topic 40: knee, swelling, oa).

Of greater interest are the latent primary care topical series which exhibit COVID-19 induced changes, but for which no return to baseline is observed over the pandemic year. Specific topics with sustained decreases in topical prevalence during the 2020 pandemic year include blood pressure screening/monitoring and cardiovascular disease screening/monitoring topics (Topic 7: bp, systolic, diastolic, htn; Topic 25: heart, bpm, systolic, diastolic; Topic 40: chest, sob, cvs). If structured EMR data, and/or billing/referral data could corroborate the finding that primary care screening/monitoring/management of cardiovascular disease may be delayed during the pandemic, this would be an important area for the primary healthcare system to consider re-prioritizing scarce resources in the near future. Persistent decreases in topical prevalence were also observed for topics encoding prescription medications (Topic 2), weight monitoring (Topic 11: weight, kg, bmi), and neurological symptoms/findings (Topic 24).

Finally, we note that many primary care topical series demonstrated stable patterns throughout the pandemic year, illustrating similar patterns as were observed 2017–2019. For example, Topic 10 and Topic 19 encode the primary healthcare system role in annual immunization programs - with peaks in topical prevalence being noted in October/November/December each year – these patterns are maintained during the pandemic year. Topics encoding specific aspects of medication management (Topic 16, Topic 17, and Topic 34) illustrate roughly flat dynamics over the study time period. Similarly, lifestyle monitoring topics (Topic 14 and Topic 33) illustrate flat topical prevalence patterns through time. We did not observe meaningful temporal variation with respect to Topic 21 representing breast health (and possibly mammography). We note that the token “breast” occurs when discussing “breast feeding”; and so the topic itself may encode a variety of breast health issues. That said, anecdotal evidence from Ontario suggests that breast cancer screening rates may be down (a meaningful trend our model has not identified). A number of topical series encoding acute health conditions are unimpacted by the pandemic, including a urinary tract infection topic (Topic 26), a vision topic (Topic 27), a foot swelling topic (Topic 29), an abdominal/bowel-movement topic (Topic 50) and a topic representing the common cold (Topic 47: nasal, sinus, congestion, nasonex). Topical series representing chronic health conditions, encoding mood/depression (Topic 32), hip osteo-arthritis (Topic 46) and blood pressure medication management (Topic 42) illustrate flat dynamics over the study timeframe. Topic 47 (dermatological issues) shows predictable harmonics through the pandemic, with expected summer highs and winter lows. Finally, topic 31 encoding familial relationships remains flat over our study time period.

In applied topic modelling communities, there does not exist consensus regarding a single methodology or metric for post-hoc determining an "optimal" topic model fitted to an empirical document collection [20,21]. In this study, we utilized a human-in-the-loop approach for determining an appropriate number of latent topical bases used to describe the primary care document collection; whereby, subject matter experts and analysts iteratively "eyeballed" learned latent topical bases, settling on a topical basis which adequately described the corpus. The human in the loop approach is popular among analysts fitting topic models to empirical datasets [22]. The approach is necessarily subjective as it does not a priori aim to choose a NMF model which optimizes some topic model validity index (e.g. topical coherence, reconstruction error, etc.). We do however perform an extensive series of sensitivity analysis to demonstrate the robustness/stability of inferences across model hyper-parameter configurations and/or stochastic data perturbations. We refit NMF temporal topic models and critically appraise latent topical time series under a number of sensitivity analysis scenarios: 1) varying NMF model complexity, 2) randomly initializing NMF latent matrices using different seeds, and 3) performing bootstrap resampling of the input DTM used in NMF temporal topic modelling. Descriptive inferences generated in this study are incredibly stable/robust, as nearly all sensitivity analyses corroborate study findings discussed above. In particular, following initiation of the COVID-19 pandemic we observe meaningful changes latent topical series encoding respiratory disease, mental health, sleep, social and familial dynamics, measurement of vitals/labs, uptake of prevention/screening maneuvers, and referrals to medical specialists. We note these inferences are rather subjective, as no quantitative metric is used to evaluate similarity of latent quantities across model fits in each independent sensitivity analysis.

This study has illustrated how temporal topic modelling applied to a large corpus of rich primary care clinical text data can identify a meaningful topical/thematic summarization which can provide policy-makers and public health stakeholders a passive, cost-effective, technology for understanding holistic impacts of the COVID-19 pandemic on the primary healthcare system and community/public-health.

4.1. Limitations and future work

In this study we apply a non-negative matrix factorization temporal topic model to a large/expressive corpus of primary care clinical notes to investigate COVID-19 pandemic effects on the physical, mental, and social health of individuals from Toronto, Canada. Our study, however, is not without limitations. Below, we speak to some general methodological limitations of our study and discuss how these might be reframed as areas for future work.

To begin, it is possible that we could have pre-processed our text data using a different computational pipeline. While we have attempted to be explicit and transparent with respect to how our final vocabulary of words/tokens was chosen, different computational pipelines could have been employed to pre-process our text corpus. For instance, differing approaches to tokenization, lemmatization, stemming, stop-word removal, and/or frequency-based word/token removal could have been implemented yielding a different analytic vocabulary. Further, rather than working with a "bag of words", we could have attempted to map documents – represented as digital character sequences – onto a finite code-set, using a pre-defined external ontology/nomenclature (e.g. ICD codes, UMLS codes, MESH codes, SNOMED codes, etc.). In this case, rather than working with collections of high dimensional term-frequency vectors, one could work with a collection of high-dimensional code-frequency vectors. A "document code matrix" could be constructed, factorized, and latent topical matrices could be generated (where topical vectors would be comprised of semantically correlated codes rather than words/tokens). It is arguable whether the resulting topical-code vectors may be more post-hoc interpretable than the topical-word vectors. Further, it may be interesting to (qualitatively) compare inferences resulting from a bag-of-words vs. a bag-of-codes topic modelling approach.

As a result of using a word/token-based approach to topic modelling, a challenge emerges with regards to how we ought to interpret the estimated topical summary vectors. Each vector represents the extent to which a collection of semantically correlated words/tokens load onto a given (probability) vector. However, human judgement/involvement is required to interpret the returned collection of topical vectors. Novel methods exist which purport to provide an automatic/algorithmic summarization of the returned topical vectors, such that subjective human-judgement can be circumvented. Different methods exist for automatic labelling of topic models. Certain methods attempt to extract representative summaries from observed text [23]. Related methods attempt to algorithmically generate representative summaries using neural language modelling techniques [24]. Alternative classes of models attempt to augment observed text data with external *meta*-data (in the form of large knowledge bases and ontologies) which can be used for automatic labelling generated topics [25,26]. Future work should continue to investigate how individuals engage with topic models and attempt to derive topical summaries which are most useful/engaging for end users (possibly via the use of automatic labelling techniques).

Finally, we note that the non-negative matrix factorization approach we employed to temporal topic model estimation is not the only/best methodology, alternative methods exist. A simple alternative might involve utilizing alternative linear algebraic models (e.g. Latent Semantic Analysis [16]) or probabilistic graphical models (e.g. Latent Dirichlet Allocation [27]) in place of our NMF model. Alternatively, we could consider more sophisticated probabilistic graphical models designed explicitly to model the dynamic evolution of topical vectors (e.g. Dynamic Topic Modelling [28]) or for which covariates can be readily included in order to model how topical content changes as a function of observed covariates (e.g. Structural Topic Modelling [29]). An altogether different, albeit applicable method, would involve representing our dataset as a ragged 3-way/3-mode tensor (i.e. row-mode = documents, column-mode = word/tokens; and depth-mode = time-slices) and using a tensor decomposition model to investigate evolution of latent topical vectors trough time (e.g. the PARAFAC2 model [30,31]). Lastly, deep learning researchers are beginning to explore techniques which jointly embed words, documents and topics in a shared space, enabling neural approaches to topic modelling (e.g. TOP2VEC [32], LDA2VEC [33], and BERTopic) which may also be useful as core modules for temporal topic modelling. Future studies which compare several of the aforementioned methods and aim to develop conclusions regarding strengths/weaknesses of the above methods for temporal topic modelling (using a combination of empirical and simulated datasets) would represent a contribution to the research community.

5. Conclusions

Primary care clinical text data contain unique and expressive information, which can be used to characterize, monitor, and evaluate a myriad of impacts the COVID-19 pandemic has had on patient/public-health. Non-negative matrix factorization temporal topic modelling of unstructured clinical text data provides a scalable and straightforward methodology for characterizing and monitoring COVID-19 pandemic impacts on the primary healthcare system and community public health. Although inferences from this unsupervised machine learning study are necessarily descriptive and hypothesis generating, the findings generated from this study are robust/stable and can provide complementary information which can be used to assist stakeholders in understanding diverse COVID-19 pandemic effects on the primary healthcare system and prioritize scarce healthcare resources as part of post-COVID-19 pandemic planning.

CRediT authorship contribution statement

Christopher Meaney: Writing – original draft, Conceptualization, Methodology, Data curation, Programming, Formal analysis. **Michael Escobar:** Writing – review & editing, Supervision, Methodology. **Rahim Moineddin:** Writing – review & editing, Supervision, Methodology. **Therese A. Stukel:** Writing – review & editing, Supervision, Methodology. **Sumeet Kalia:** Writing – review & editing, Methodology, Data curation. **Babak Aliarzadeh:** Writing – review & editing, Methodology, Data curation. **Tao Chen:** Writing – review & editing, Methodology, Data curation. **Braden O'Neill:** Writing – review & editing, Methodology. **Michelle Greiver:** Writing – review & editing, Supervision, Methodology, Investigation.

Funding statement

The study was supported by funding provided by a Foundation-Grant (FDN 143303) from the Canadian Institutes of Health Research (CIHR). The funding agency had no role in the study design, the collection, analysis, or interpretation of data, the writing of the report, or the decision to submit the report for publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2022.104034.

References

[1] W. Wiersinga, A. Rhodes, Cheng, et al., Pathophysiology, Transmission, Disgnosis and Treatment of Coronavirus Disease 2019 (COVID-19): A Review, Journal of the American Medicine Association 324 (8) (2020) 782–793.

[2] A. Nalbandian, K. Sehgal, A. Gupta, M.V. Madhavan, C. McGroder, J.S. Stevens, J. R. Cook, A.S. Nordvig, D. Shalev, T.S. Sehrawat, N. Ahluwalia, B. Bikdeli, D. Dietz, C. Der-Nigoghossian, N. Liyanage-Don, G.F. Rosner, E.J. Bernstein, S. Mohan, A. A. Beckley, D.S. Seres, T.K. Choueiri, N. Uriel, J.C. Ausiello, D. Accili, D. E. Freedberg, M. Baldwin, A. Schwartz, D. Brodie, C.K. Garcia, M.S.V. Elkind, J. M. Connors, J.P. Bilezikian, D.W. Landry, E.Y. Wan, Post-Acute COVID-19 Syndrome, Nat. Med. 27 (4) (2021) 601–615.

[3] J. He, Y. Guo, R. Mao, J. Zhang, Proportion of Asymptomatic Coronavirus Disease 2019: A Systematic Review and Meta-Analysis, J. Med. Virol. 93 (2) (2021) 820–830.

[4] E.A. Meyerowitz, A. Richterman, I.I. Bogoch, N. Low, M. Cevik, Towards an accurate and systematic characterisation of persistently asymptomatic infection with SARS-CoV-2, Lancet Infectious Disease 21 (6) (2021) e163–e169.

[5] National Library of Medicine. COVID-19 Scope Note. Retrieved May 27, 2021 from the following URL: https://meshb.nlm.nih.gov/record/ui?ui=D000086382.

[6] A. Docherty, E. Harrison, C. Green, et al., Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study, British Medical Journal 369 (2020) M1985–M2014.

[7] N. Imai, K.A.M. Gaythorpe, S. Abbott, S. Bhatia, S. van Elsland, K. Prem, Y. Liu, N. M. Ferguson, Adoption and Impact of Non-Pharmaceutical Interventions for COVID-19, Welcome Open Research 5 (2020) 59, https://doi.org/10.12688/wellcomeopenres10.12688/wellcomeopenres.15808.1.

[8] S. Lai, N. Ruktanonchai, L. Zhou, et al., Effect of Non-Pharmacetutical Interventions to Contain COVID-19 in China, Nature 585 (7825) (2020) 410–413.

[9] M. Douglas, S. Katikireddi, G. McCartney, Mitigating the Wider Health Effects of COVID-19 Pandemic Response, Br. Med. J. 369 (2020) M1557–M1572.

[10] Wikipedia1. List of Largest Cities in North America by Population Size. Retrieved May 27, 2021 from URL: https://en.wikipedia.org/wiki/List_of_North_American_cities_by_population.

[11] Toronto Public Health COVID Tracker. Retrieved May 27, 2021 from the following URL: https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto/.

[12] P. Turney, P. Pantel, From Frequency to Meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research 37 (2010) 141–188.

[13] C. Manning, P. Raghavan, H. Shutze, An Introduction to Information Retrieval, Cambridge University Press, 2009.

[14] D. Lee, S. Seung, Learning the Parts of an Object by Non-Negative Matrix Factorization, Nature 401 (1999) 788–791.

[15] D. Lee, S. Seung, Algorithms for Non-Negative Matrix Factorization, Advances in Neural Information Processing Systems (2001) 556–562.

[16] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science 41 (6) (1990) 391–407.

[17] T. Griffiths, M. Steyvers, Probabilistic Topic Models. 2007; In Handbook of Latent Semantic Analysis. Chapter 21.

[18] M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, R.J. Plemmons, Algorithms and Applications for Approximate Non-Negative Matrix Factorization, Comput. Stat. Data Anal. 52 (1) (2007) 155–173.

[19] M. Udell, C. Horn, R. Zadeh, S. Boyd, Generalized Low Rank Models, Foundations and Trends in Machine Learning 9 (1) (2016) 1–118.

[20] J. Chang, S. Gerrish, W. Chong, J. Boyd-Graber, D. Blei, Reading Tea Leaves: How Humans Interpret Topic Models. Proceedings of Neural Information Processing Systems (2009).

[21] C. Doogan, W. Buntine, Topic Model or Topic Twaddle? Re-Evaluating Semantic Interpretability Measures, NAACL. (2021) 3824–3848.

[22] P. Matthews, Human In the Loop Topic Modelling, International Society for Knowledge Organization. (2019) 1–31.

[23] X. Wan, T. Wang, Automatic Labelling of Topic Models Using Text Summaries, ACL. (2016) 2297–2305.

[24] A. Alokaili, N. Aletras, M. Stevenson, Automatic Generation of Topic Labels, ACM. (2020) 1965–1968.

[25] I. Hulpus, C. Hayes, D. Greene, Unsupervised Graph Based Topic Labelling Using DBPedia, ACM. (2013) 465–474.

[26] M. Allahyari, S. Pouriyeh, K. Kochut, H.R. Arabnia, A knowledge-based topic modeling approach for automatic topic labeling, International Journal of Advanced Computer Science and Applications. 8 (9) (2017) 335.

[27] D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research. 3 (2003) 993–1022.

[28] D. Blei, J. Lafferty, Dynamic Topic Models, Proceedings of ICML. (2006) 113–120.

[29] M. Roberts, B. Stewart, E. Airoldi, A Model of Text for Experimentation in the Social Sciences, J. Am. Stat. Assoc. 111 (515) (2016) 988–1003.

[30] A. Cichocki, R. Zdunek, A.H. Phan, S. Amari, Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation, Wiley, 2009.

[31] P.M. Kroonenberg (Ed.), Wiley Series in Probability and StatisticsApplied Multiway Data Analysis, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2008.

[32] D. Angelov, TOP2VEC: Distributed Representations of Topics, Arxiv. (2020) 1–25.

[33] C. Moody, Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec, Arxiv. (2016) 1–8.