The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3165–3184
https://doi.org/10.1214/24-AOAS1930

DYNAMIC TOPIC LANGUAGE MODEL ON HETEROGENEOUS CHILDREN'S MENTAL HEALTH CLINICAL NOTES

BY HANWEN YE[1,a], TATIANA MORENO[2,c], ADRIANNE ALPERN[2,d], LOUIS EHWERHEMUEPHA[2,e] AND ANNIE QU[1,b]

[1]*Department of Statistics, University of California,* [a]*hanweny@uci.edu,* [b]*aqu2@uci.edu*
[2]*Children's Hospital of Orange County,* [c]*Tatiana.Moreno@choc.org,* [d]*AAlpern@choc.org,* [e]*LEhwerhemuepha@choc.org*

Mental health diseases which affect children's lives and well-beings have received increased attention since the COVID-19 pandemic. Analyzing psychiatric clinical notes with topic models is critical to evaluating children's mental status over time. However, few topic models are built for longitudinal settings, and most existing approaches fail to capture temporal trajectories for each document. To address these challenges, we develop a dynamic topic model with consistent topics and individualized temporal dependencies on the evolving document metadata. Our model preserves the semantic meaning of discovered topics over time and incorporates heterogeneity among documents. In particular, when documents can be categorized, we propose a classifier-free approach to maximize topic heterogeneity across different document groups. We also present an efficient variational optimization procedure adapted for the multistage longitudinal setting. In this case study, we apply our method to the psychiatric clinical notes from a large tertiary pediatric hospital in Southern California and achieve a 38% increase in the overall coherence of extracted topics. Our real data analysis reveals that children tend to express more negative emotions during state shutdowns and more positive when schools reopen. Furthermore, it suggests that sexual and gender minority (SGM) children display more pronounced reactions to major COVID-19 events and a greater sensitivity to vaccine-related news than non-SGM children. This study examines children's mental health progression during the pandemic and offers clinicians valuable insights to recognize disparities in children's mental health related to their sexual and gender identities.

1. Introduction.

1.1. *Motivation*. Mental health conditions, such as anxiety, depression, and substance abuse, are prevalent among children and can have long-lasting impacts on their social relationships and academic performance. Without proper intervention, mental health conditions can lead to school absences, academic failure, isolation from peers, and—in some cases—an increased risk of suicide. Unfortunately, the outbreak of COVID-19 has intensified pediatric mental health issues due to the extent level of enforced physical/social isolation (Wu et al. (2021), Ravens-Sieberer et al. (2022)). In particular, sexual and gender minority (SGM) youth who live in low-supportive home environments and lose access to previously affirming environments, such as schools and social activities, are placed at an increased risk for abuse and rejections from their family members (McGeough and Sterzing (2018), Thoma et al. (2021)). Therefore, to develop effective interventions and create post-pandemic support systems, it is crucial for mental health professionals to understand the dynamic changes and disparities in children's mental status concerning their sexual and gender identities during the pandemic (Salerno et al. (2020)).

Received December 2023; revised April 2024.
Key words and phrases. Classifier-free, multistage topic language models, sexual and gender identity, time-consistent topics, variational inference.

However, evaluating children's mental health, in general, is challenging due to the self-reported symptoms and complex heterogeneity among subjects. For instance, studies found that girls and SGM children tend to report their symptoms differently and are more likely to be diagnosed with depression, compared to boys and non-SGM youth, even though they express similar symptoms (Afifi (2007), Rosenfield and Mouzon (2013), Marshal et al. (2011), Plöderl and Tremblay (2015), Russell and Fish (2016)). To mitigate this bias, questionnaires and telephone surveys with validated rating scale (Penninx et al. (2008), Barry (2014), Boyd et al. (2013)) are conducted to quantify patients' mental health symptoms with derived metrics. Yet these study designs still suffer from selection bias and are unable to capture important life events or stressors impacting the patient's mental health.

In this study we leverage inpatient mental health unit notes from a large tertiary pediatric hospital in Southern California to evaluate children's mental health over the pandemic period. Compared to survey metrics, clinical notes contain a more contextualized background of a patient (e.g., mental health history, hospitalization reasons, interactions with clinicians, etc). Our goal is to identify major life events and stressors from these clinical records and track their dynamic shifts to quantify mental health status changes among nearly 2600 inpatient children throughout the pandemic. Importantly, these contextual factors play a crucial role in revealing the prevalent and long-lasting themes in mental health, such as depression, anxiety disorders, and suicidal intentions (Scott (1958), Ronald et al. (2010), National Institute of Mental Health (2021), Ciechanowski, Jemielniak and Silczuk (2023)). By understanding the evolution trends of these themes, we offer valuable insights into the progression of children's mental health throughout the pandemic era. Moreover, taking into account each child's unique life experience and reactions, we incorporate individual-level heterogeneity and aim to uncover distinct trends between SGM and non-SGM children. This allows investigators to better recognize mental health disparity among SGM youth during COVID-19 and develop tailored post-pandemic support systems concerning sexual and gender identities.

1.2. *Literature review.* Unsupervised topic modeling is a popular statistical approach which aligns well with our objective of discovering abstract themes (i.e., topics) from a large corpus of text data. By clustering common patterns and keywords across multiple documents, topic models uncover the underlying semantics associations and summarize lengthy documents into a manageable number of interpretable topics. In the existing literature, topic modeling can be typically categorized into two major frameworks: Bayesian probabilistic topic models (BPTMs) and neural topic models (NTMs). A BPTM proposes a probabilistic generative model of a document and applies Bayesian inference procedures to estimate the posterior of latent topics. Representative methods include latent Dirichlet allocation (LDA) (Blei, Ng and Jordan (2003), Blei and Lafferty (2006a)), the dynamic LDA (Blei and Lafferty (2006b), Wang, Blei and Heckerman (2012)), where topics evolve with time, and the supervised LDA (Mcauliffe and Blei (2007), Roberts et al. (2014), Li, Ouyang and Zhou (2015), Sridhar, Daumé III and Blei (2022)) with augmentation of document metadata. However, BPTMs generally suffer from low sampling efficiency and high technical difficulties in customizing the optimization procedure for each model prior specification. NTMs, on the other hand, are based on neural networks to model the relationships between words and topics. Benefiting from the standard gradient descent optimization procedure, NTMs can be easily integrated into different application cases and achieve high training efficiency on large datasets. Under the NTMs framework, one can find topics via clustering the word embedding representations (Thompson and Mimno (2020), Sharifian-Attar et al. (2022)) or modeling topics as latent variables in autoregressive models (Larochelle and Lauly (2012), Gupta et al. (2019)), generative adversarial networks (GANs) (Wang, Zhou and He (2019), Hu et al. (2020)). In particular, the variational autoencoder (VAE)-based NTMs (Miao, Yu and Blunsom (2016),

Srivastava and Sutton (2017), Lin, Hu and Guo (2019)) have received the most attention due to their ability to capture complex word-topic associations with deep learning architecture and, meanwhile, provide probabilistic interpretations to the latent topics based on variational inference.

Among these frameworks discussed above, three methods offer promising solutions to our problem: the multistage dynamic LDA (Blei and Lafferty (2006b)), the supervised LDA (Mcauliffe and Blei (2007)), and SCHOLAR, which is the VAE-NTM with metadata augmentation (Card, Tan and Smith (2017)). However, none of these methods is directly applicable to our specific use case. First, the time-varying topics found by the dynamic LDAs may distort the meaning of each topic, fail to capture consistent mental health themes, and impose difficulties in interpreting the topic proportion trend due to the loss of time consistency. In addition, both supervised LDAs and SCHOLAR are single-stage topic models and rely on classifiers, instead of topic distributions, to differentiate groups. Moreover, few studies have extended the VAE-NTMs framework to the multistage longitudinal setting, though VAEs with spatiotemporal dependencies have been actively explored in the computer vision field (Gulrajani et al. (2016), Casale et al. (2018), Fortuin et al. (2020), Ramchandran et al. (2021)).

1.3. *Contribution*. This paper proposes a novel multistage dynamic VAE-NTM, namely, Heterogeneous Classifier-Free Dynamic Topic Model (HCF-DTM), with grouping information to address the challenges discussed above. In contrast to the dynamic LDAs, our method finds a number of time-consistent topics among all documents at any time point. This not only maintains the semantic meaning of discovered topics over the investigation period but also enables the direct use of obtained topic proportions to infer the dynamic change in the popularity of each topic. Additionally, we augment the document metadata into the topic-finding procedure to account for the longitudinal heterogeneity among documents. Moreover, compared to the supervised LDAs, our proposed model increases the groupwise differences directly on the latent topic distribution level. Instead of relying on additional downstream classifiers, we introduce the counterfactual topic proportions and maximize the interdistributional distances between topic proportions of the ground truth group and those as if the documents belonged to the other groups. As a result, the distinct characteristics of each group and their corresponding topic evolution trend can be easily identified.

The main clinical contributions of our paper are as follows. First, this work is among the earliest to utilize topic models on unstructured psychiatric clinical notes to unfold the longitudinal mental health disparities concerning sexual and gender identities during the pandemic. Knowledge of possible pronounced reactions among SGM children toward major COVID-19 events advocates for developing tailored post-pandemic treatments. Clinicians can design programs, such as virtual support groups and online mental health platforms (Whaibeh, Vogt and Mahmoud (2022), Karim et al. (2022), McGregor et al. (2023)), to help address the heightened stress, anxiety, and isolation experienced by SGM youth when their previously accessible resources become limited due to the pandemic. In addition, our study assists clinicians in further examining SGM-related contextual stressors. This not only raises community awareness about the unique challenges faced by SGM children, promoting family education and fostering community support, but also informs future research on the mental health of SGM children and gets better prepared for future pandemics or social crises.

The remainder of this article is structured as follows. In Section 2 we introduce the notations and limitations of the dynamic LDA method. In Section 3 we propose the generative process of HCF-DTM, present variational inference details, and introduce a classifier-free approach to maximize heterogeneity between groups. Section 4 explains the implementation algorithm. In Section 5 extensive simulation results are presented to illustrate the performance

advantages of our proposed HCF-DTM method. In Section 6 we apply the proposed method to the psychiatric inpatient clinical notes provided by a large tertiary pediatric hospital in Southern California. Lastly, we conclude with discussions in Section 7. Technical details and proofs are provided in the Supplementary Material (Ye et al. (2024)).

2. Background and related works.

2.1. *Notation and preliminary.* Consider a balanced multistage study where N participants undergo a total T finite number of stages (visits). Each participant belongs to a corresponding group. For the illustration purpose, we consider a two-group scenario (e.g., SGM and non-SGM), denoted as $Y_i \in \mathcal{Y} = \{0, 1\}$. At the tth stage, where $1 \le t \le T$, a set of subject's time-varying covariates and a clinical note (document) are recorded. The structured subjects' covariates are regarded as the metadata, denoted as $X_{it} \in \mathcal{X}_t$, and the unstructured clinical notes are the text data of interest.

To formalize the unstructured text data, we assume that the number of unique words across all recorded documents is V (vocabulary size). By assigning a unique ID to each word, we represent any document of an arbitrary number of words N_d with a vector of constant size V, that is, $(\text{cnt}_{w_1}, \text{cnt}_{w_2} ..., \text{cnt}_{w_v}) \in \mathbb{Z}^V_{\ge 0}$. This vector is known as the Bag of Words (BOW) representation, where each element counts the number of appearances of the corresponding word in a document. With BOW we are able to vectorize the entire corpus of documents over T time points with a structured tensor of size $T \times N \times V$, denoted as $\mathcal{D} = [\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T]$. Now, suppose at each time point t that there exist K topics within documents $\mathcal{D}_t$, and each document is a mixture of these topics. To capture the proportions of each topic found in a document, we create a document-topic matrix $\Theta_{t,N\times K}$. Furthermore, we define a word-topic matrix $\beta_{t,V\times K}$ to represent the relevance of each word to the K topics. The primary goal of topic modeling is to find the matrices $\Theta_{t,N\times K}$ and $\beta_{t,V\times K}$ which can best represent the documents $\mathcal{D}_t$ at time stage t (i.e., $\mathcal{D}_{t,N\times V} \approx \Theta_{t,N\times K} \cdot \beta^{\mathsf{T}}_{t,V\times K}$).

2.2. *Multistage dynamic LDA.* Traditional nonnegative matrix factorization (NMF) methods (Paatero and Tapper (1994), Lee and Seung (1999), Li et al. (2021)) view this problem as a matrix decomposition task, where $\Theta_{t,N\times K}$ and $\beta_{t,V\times K}$ are treated as two lower-rank matrices with $1 \le K \ll \min(N, V)$. By minimizing the distance (e.g., Frobenius norm) between $\mathcal{D}_{t,N\times V}$ and $\Theta_{t,N\times K} \cdot \beta^{\mathsf{T}}_{t,V\times K}$, the NMF methods estimate the representative topics and proportions. However, NMF is not a probabilistic method and, therefore, is limited in providing valid statistical inferences to the estimated topics. LDAs, on the other hand, provide probabilistic distributions on $\Theta_{t,N\times K}$ and $\beta_{t,V\times K}$ and incorporate them into the generative process of a document. For example, in the multistage dynamic LDA (Blei and Lafferty (2006b)) generative process 1 defined as below, δ^2, ξ^2, and a^2 are the variance priors for the latent topics, and $\sigma(x_{1:V})_j = \frac{\exp x_j}{\sum_{j=1}^V \exp x_j}$ is a softmax function constraining the unbounded multivariate normal mean parameters to a valid multinomial probability simplex.

Based on the Generative Process 1 formulation, we identify the following three limitations of the dynamic LDA to our application. First, the word-topics distribution β_t, which characterizes the semantic meaning of each topic, changes at each time stage. Commonly, the meaning of topics can be summarized by a broader theme, and the magnitude of change can be controlled via the variance prior δ^2. However, finding a suitable prior requires manual inspections of the word-clouds at each time stage to interpret and confirm that the topics are under the same theme. As the number of stages and topics increases, the cumbersome nature of this process will inevitably impose a significant challenge in topic interpretations and representations. Second, the topic proportion $\theta_{t,d,1:K}$ is not directly correlated to its precursor $\theta_{t-1,d,1:K}$ which describes the topic proportions of the same document d from the previous time stage. Instead, all $\{\theta_{t,d,1:K}\}_{d=1}^N$ share a common corpus-level hyperparameter α_t, mak-

Generative Process 1 Dynamic LDA

1. Draw word-topics distribution: $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \delta^2 I)$.
2. Draw document-topics proportion: $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \xi^2 I)$.
3. For each document d at time t:
   a. Draw topics proportion: $\eta_{t,d,1:K} \sim \mathcal{N}(\alpha_t, a^2 I)$, $\quad \theta_{t,d,1:K} = \sigma(\eta_{t,d,1:K})$.
   b. For each word at position j:
      (i). Draw a topic for this word: $Z_{t,d,j} \sim \text{Mult}(\theta_{t,d,1:K})$.
      (ii). Draw a word: $W_{t,d,j} \sim \text{Mult}(\sigma(\beta_{t,1:V,Z_{t,d,j}}))$.

ing it difficult for the current process to account for document-level heterogeneity. Lastly, the generative process above does not incorporate subjects' metadata and group information. Despite that the follow-up supervised LDA (Mcauliffe and Blei (2007)) accounts for the group information by adding an extra classification task to the end of the generative process (i.e., $Y_d | Z_{t,d,.}, \phi, \gamma \sim \text{logit-Normal}(\phi^\intercal \cdot Z_{t,d,.}, \gamma^2)$), the performance of topic separation relies heavily on the prior parameters of the classifier, ϕ and γ, and still, it is challenging to distinguish whether a strong classification result is due to the quality of the classifier or the actual presence of interpretable and separable underlying topic distributions.

3. Methodology. To address the challenges listed above, we propose the Heterogeneous Classifier-Free Dynamic Topic Model (HCF-DTM) with consistent topic interpretations and dynamic incorporation of documents' time-varying metadata. In the following we describe the generative process and a detailed variational inference procedure of our proposed model. In addition, we introduce a novel classifier-free approach to directly maximize the groupwise heterogeneity among topics with a notion of counterfactual topic distributions.

3.1. *Heterogeneous DTM with consistent topics.* The main objective of our proposed model is to identify a set of time-consistent topics while accounting for the evolving heterogeneity within documents over a specified longitudinal timeframe. This subsection presents the detailed specification of HCF-DTM in Generative Process 2 and illustrates with the corresponding graphical model in Figure 1.

Compared to the Generative Process 1 of previous DTMs, our method differentiates itself in three significant ways. First, instead of allowing each time point to have its own word-topic matrices $\{\beta_t\}_{t=1}^T$, we remove the time dependency and assume the existence of a single word-topic matrix, β, which is shared by all documents and held constant regardless of time or group memberships. Specifically, our approach achieves this by parameterizing the generative distribution of β with a single time-invariant mean prior β_0 and sampling it once at the beginning of the process. In addition, we keep the generative variance prior, $\delta^2 I$, diagonal to ensure orthogonality and maximize disparities among topics. As a result, the topic matrix β provides a consistent and distinguishable interpretation for each topic at every time

Generative Process 2 Heterogeneous DTM with consistent topics

1. Draw time-consistent word-topics distribution: $\beta \sim \mathcal{N}(\beta_0, \delta^2 I)$.
2. For each document d at time t:
   a. Draw topic proportions:
      $\eta_{t,d,1:K} \mid \eta_{t-1,d,1:K}, X_{d,t}, Y_d, \phi_t \sim \mathcal{N}(f_{t,\phi_t}(\eta_{t-1,d,1:K}, X_{d,t}, Y_d), a^2 I)$,
      $\theta_{t,d,1:K} = \sigma(\eta_{t,d,1:K})$.
   b. For each word at position j:
      (i). Draw the word: $W_{t,d,j} \sim \text{Mult}(\theta_{t,d,1:K} \cdot \sigma(\beta)^\intercal)$.

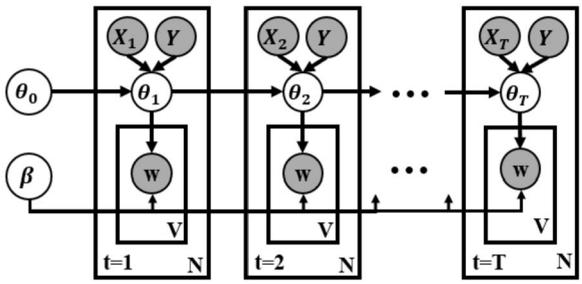

FIG. 1. *A graphical model of the heterogeneous DTM in a balanced multistage longitudinal setting, where θ_0 is a prior for the initial topic proportions. Topics β's are held constant and provided at every time stage. Both θ and β are the latent variables, whereas W, X, and Y are observed variables.*

point. This is more desirable, as it helps uncover long-lasting topics and guarantees consistent themes without the need to navigate through a multitude of topics $\{\beta_t\}_{t=1}^T$ over time.

Apart from providing consistent topics, we augment documents' metadata into the generative process and establish individualized trajectories of the longitudinal topic proportions. Notably, since the topics β do not vary with time in our design, the topic proportions $\{\theta_{t,d,1:K}\}_{t=1}^T$ encapsulate all temporal correlations within each document d. In particular, we utilize a function $f_t : \mathbb{R}^K \times \mathcal{X}_t \times \mathcal{Y} \mapsto \mathbb{R}^K$ parameterized by prior ϕ_t to capture the mean trend of the topic proportions. At every time stage t for a document d, the mean function f_t takes three inputs: the past topic proportions $\theta_{t-1,d,1:K}$, subject's longitudinal covariates $X_{d,t}$, and group membership Y_d to provide predictions for the expected topic proportions. Consequently, the topic proportions Θ_t incorporate the dynamic heterogeneity among subjects and are able to control for confounders, which may affect their distributions. More importantly, we let the topic proportions Θ_t directly depend on Θ_{t-1} from the previous stage rather than a corpus-level hyperparameter α_t used in dynamic LDA. This enables us to better track the individual progression of topic proportions on the document/subject level.

Furthermore, our heterogeneous DTM generative process simplifies the optimization procedure. By estimating the time-consistent latent topics simultaneously across different time periods, we reduce the number of model parameters and decrease the inference complexity level. Additionally, inspired by the work of Miao, Yu and Blunsom (2016), we collapse the latent random variable Z, which determines the topic assignment for each word, from the word sampling distribution based on the following results:

$$p(w_t|\theta_t, \beta) = \sum_{z=1}^{K} p(z|\theta_t) \cdot p(w_t|z, \beta) = \sum_{z=1}^{K} \theta_{t,z} \cdot \sigma(\beta_z) = \left(\theta_t \cdot \sigma(\beta)^\intercal\right)_w. \tag{1}$$

Unlike the two-layer hierarchical dependency outlined in Generative Process 1, we model the word generative process jointly with the topics β and their proportions Θ_t, as in $W_t \sim \text{Mult}(1, \theta_t \cdot \sigma(\beta)^\intercal)$. This collapsing approach has the advantage of eliminating the need to sample the discrete random variable Z during the optimization process, which is considered to be more challenging.

Finally, we aim to estimate the posterior of the latent topic proportions via the Bayes rule according to the defined prior distributions of the latent topics and the likelihood model of the documents in Generative Process 2. To simplify notations, we use θ_t, X_t, and Y as shorthand to represent $\theta_{t,d,1:K}$, $X_{d,t}$, and Y_d, respectively, for each document d from this point onward,

$$p(\theta_t \mid \theta_{1:(t-1)}, \boldsymbol{w}_t, X_t, Y, \beta) = \frac{p(\theta_t|\theta_{1:(t-1)}, X_t, Y) \cdot p(\boldsymbol{w}_t|\theta_t, \beta)}{\int_{\theta_t} p(\boldsymbol{w}_t|\theta_t, \beta) \cdot p(\theta_t|\theta_{1:(t-1)}, X_t, Y)\, d\theta_t}. \tag{2}$$

With the posterior in equation (2), we can infer the latent topic proportions θ_t at any time t from the observed document data. However, obtaining this posterior distribution is extremely

difficult. This is because the marginalized data distribution $p(\boldsymbol{w}_t|\theta_{1:(t-1)}, \beta)$ in the denominator is intractable due to the unbounded space of the latent variable θ_t during integration. To address this challenge, we leverage the variational Bayes and propose a novel solution to extend the inference procedure of heterogeneous topic models to a multistage longitudinal setting. A detailed explanation of the longitudinal variational Bayes approach under our generative process specifications is provided in the next subsection.

3.2. *Longitudinal variational inference.* Variational inference is a powerful technique for estimating the intractable posterior of interest P with a parametrizable distribution Q, such as a Gaussian distribution. To ensure Q is a valid approximation to P, variational methods minimize the Kullback–Leibler (KL) divergence between the two distributions. This transforms the inference problem into an efficient optimization task, as future posterior inference can be conducted directly from Q. In our specific case, we aim to find the optimal set of parameters $\psi^*_{1:T}$ over the T number of stages so that the parameterized approximating distribution $Q^* \doteq Q_{\psi^*_{1:T}}$ is as close as possible to the true topic posterior distribution P, that is,

$$\psi^*_{1:T} = \underset{\psi_{1:T}\in\Psi^{|T|}}{\operatorname{argmin}} \mathbb{KL}\big(Q_{\psi_{1:T}}(\theta_{1:T}|\boldsymbol{w}_{1:T}, X_{1:T}, Y)||P(\theta_{1:T}|\boldsymbol{w}_{1:T}, X_{1:T}, Y, \beta)\big) \tag{3}$$

$$= \underset{\psi_{1:T}\in\Psi^{|T|}}{\operatorname{argmin}} \underbrace{\mathbb{KL}\big(Q_{\psi_{1:T}}(\theta_{1:T}|\boldsymbol{w}_{1:T}, X_{1:T}, Y)||p(\theta_{1:T}|X_{1:T}, Y)\big)}_{\textbf{Approximation error}: \text{Variational distribution } Q \text{ and geneartive prior } p} \tag{4}$$

$$\underbrace{-\mathbb{E}_{\theta_{1:T}\sim Q_{\psi_{1:T}}}\big(\log P(\boldsymbol{w}_{1:T}|\theta_{1:T}, \beta)\big)}_{\textbf{Reconstruction error}: \text{Model likelihood from variational topics}}.$$

In a single-stage setting, optimizing ψ^* is straightforward through minimizing the negative evidence lower bound (ELBO) in equation (4). However, the inference is more difficult to define when there are multiple stages involved, since additional temporal dependencies arise among the topic proportions $\theta_{1:T}$. Moreover, as $\theta_{1:T}$ are latent variables and not directly observable, we sample θ_t along with its temporal predecessors $\theta_{1:(t-1)}$ at each time stage, which increases the computational complexity exponentially with the number of stages. Therefore, it is more desirable to derive a variational objective, which can efficiently break the long dependent sequence into smaller sets of stages while preserving temporal correlations in a multistage longitudinal setting. Before presenting our longitudinal ELBO, we first introduce the following regularity assumptions.

ASSUMPTION 1 (Markov property of latent topic proportions). Topic proportions at the next stage only depend on the current stage, not on any past stages: $\theta_{t+1} \perp\!\!\!\perp \theta_1, \ldots, \theta_{t-1}|\theta_t$.

ASSUMPTION 2 (Independence generation of documents). The distribution of every word from documents at time t only depends on the time-consistent topics and current-stage topic proportions: $\boldsymbol{w}_t \perp\!\!\!\perp (\boldsymbol{w}_{1:(t-1)}, \theta_{1:(t-1)}, X_{1:t}, Y) \mid \theta_t, \beta$.

Assumption 1 relaxes the temporal dependencies between topic proportions which are more than two stages apart, since topics of a document are majorly influenced by the ones from the previous stage. In other words, current-stage topic proportions capture all the past information required for the proportions at the next stage. Assumption 2 states that it is sufficient to generate all words in a document given the current-stage topic proportions and topics, as subjects' heterogeneity and potential confounders have been considered in θ_t by the mean trend function f_t. Under these two assumptions and the variational distribution factorization $Q_{\psi_{1:T}}(\theta_{1:T}, \boldsymbol{w}_{1:T}, X_{1:T}, Y) = \prod_{t=1}^{T} q_{\psi_t}(\theta_t, \boldsymbol{w}_t, X_t, Y)$, we present the following ELBO as the variational objective under our multistage heterogeneous DTM Generative Process 2.

PROPOSITION 1. *Under Assumptions 1–2, the evidence lower bound (ELBO) for a single document generated by Process 2 over a finite T-stage longitudinal time horizon is*

$$\log P(\boldsymbol{w}_{1:T}|X_{1:T}, Y, \beta) \tag{5}$$

$$\geq \underbrace{-\mathbb{KL}(q_{\psi_1}(\theta_1|\boldsymbol{w}_1, X_1, Y)||p(\theta_1|\theta_0, X_1, Y)) + \mathop{\mathbb{E}}_{\theta_1\sim q_{\psi_1}}(\log P(\boldsymbol{w}_1|\theta_1, \beta))}_{\textit{Single-stage ELBO for the first stage}} \tag{6}$$

$$+ \underbrace{\sum_{t=2}^{T}\{-\mathop{\mathbb{E}}_{\theta_{t-1}\sim q_{\psi_{t-1}}}(\mathbb{KL}[q_{\psi_t}(\theta_t|\boldsymbol{w}_t, X_t, Y)||p(\theta_t|\theta_{t-1}, X_t, Y)]) + \mathop{\mathbb{E}}_{\theta_t\sim q_{\psi_t}}(\log P(\boldsymbol{w}_t|\theta_t, \beta))\}}_{\textit{Temporal-dependent ELBO for the follow-up stages}}.$$

Equation (6) provides a lower bound for the log-likelihood of a document generated according to Generative Process 2. Maximizing this lower bound is equivalent to minimizing the negative ELBO in equation (3), where both optimizations result in the same set of optimal variational parameters $\psi^*_{1:T}$. Under the regularity assumptions, the presented longitudinal ELBO can be further decomposed into two components: the standard single-stage ELBO for the first stage, and the temporal-correlated ELBO for the follow-up stages. Specifically, the latter divides all future stages into adjacent-stage pairs based on Assumption 1. By calculating the KL divergence of current-stage topic proportions using the proportions sampled from the variational distribution at the previous stage, the approximation error across all time stages can be divided into adjacent stages. This leads to a more efficient inference process, as the adjacent-stage dependency eliminates the need for sampling proportions of the entire time sequence at each stage. Additionally, based on Assumption 2, the reconstruction error can be calculated individually at each time stage, even without the pairwise dependency, which further enhances the efficiency of the variational learning process.

In summary, the longitudinal ELBO shown in equation (6) extends variational inference for DTMs to a multistage longitudinal setting. It provides an efficient optimization objective to approximate the true posterior of the topic proportions according to our proposed heterogeneous DTMs generative process. However, as the posterior of topic proportions θ_t depends on the topics β shown in equation (2), the proportion of a topic can vary significantly based on the learned latent topics. For instance, the differences in the topic proportion distributions between non-SGM and SGM youth may be more pronounced if the topic is related to mental health rather than physiological measures. Depending on the provided latent topics, the estimated topic proportions may not optimally present the heterogeneity from the groups. Thus, to better understand groupwise differences, we propose a classifier-free approach via distributional distances to learn the latent topics.

3.3. *Groupwise topic separation.* In this subsection our goal is not only to identify the most representative topics under Generative Process 2 but, more importantly, to maximize the heterogeneity between groups in their respective proportions. To achieve this, we introduce the counterfactual topic distributions, which describe the topic proportion distributions for subjects who have the same documents $w_{1:T}$ and measurements $X_{1:T}$ but belong to different groups Y. The value function of HCF-DTM is presented as follows:

$$V^{\mathrm{HCF}}(\psi_{1:T}) = \underbrace{-\mathbb{KL}\big(Q_{\psi_{1:T}}(\theta_{1:T}|\boldsymbol{w}_{1:T}, X_{1:T}, Y)||P(\theta_{1:T}|\boldsymbol{w}_{1:T}, X_{1:T}, Y, \beta)\big)}_{\text{Variational inference objective (6)}} \tag{7}$$

$$+ \underbrace{\mathrm{Dist}\big(\{Q_{\psi_{1:T}}(\theta_{1:T}|\boldsymbol{w}_{1:T}, X_{1:T}, Y = y)\}_{y\in\mathcal{Y}}\big)}_{\text{Groupwise topic proportion distribution distance}},$$

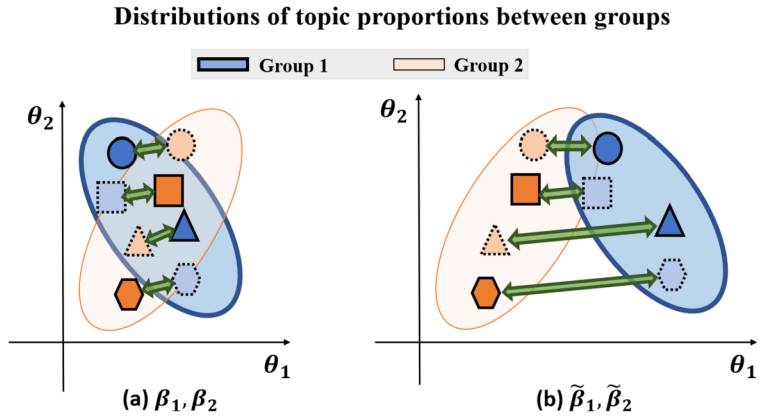

FIG. 2. *Comparison of topic proportion distributions under two sets of latent topics*: β_1, β_2 *vs.* $\tilde{\beta}_1$, $\tilde{\beta}_2$. *The shapes displayed on a two-dimensional plane correspond to optimal topic proportions for each latent topic specification. Solid outlines indicate proportions from the true group identity, while dashed outlines represent counterfactual proportions.*

where information radius (Sibson (1969/70)), average divergence score (Sgarro (1981)), or mutual information (MI) in a two-group setting can be selected as the distance metrics. To represent the counterfactual distribution, we use the constructed variational Q by changing its group membership covariate Y under the *no unmeasured confounding* assumption (Robins (1986)). The main objective is to find the set of variational parameters which maximizes the value function, that is, $\psi^*_{1:T} = \text{argmax}_{\psi_{1:T} \in \psi^{|T|}} V^{\text{HCF}}(\psi_{1:T})$. In particular, the first term encourages closer approximation of variational distribution Q to the underlying posterior P; meanwhile, the second term aims to increase the distances among the marginal topic proportion distributions, $P(\theta_{1:T}|Y)$, for each group membership Y.

The proposed distance maximizing approach has the following two advantages. First, it provides explicit guidance to learn the latent topics that have the largest groupwise difference in their corresponding proportions. Due to the co-dependency between the latent topics and their proportions, the maximum amount of group heterogeneity, which can be captured by the topic proportions, depends on the provided latent topics. To illustrate this, consider the example of a two-group scenario shown in Figure 2. Compared to latent topics β_1 and β_2, topics $\tilde{\beta}_1$ and $\tilde{\beta}_2$ lead to a larger averaged distance between the topic proportion marginal distributions and, noticeably, can better disentangle the group identities. To make $\tilde{\beta}_1$ and $\tilde{\beta}_2$ more likely to be identified in practice, the proposed second term in value function (7) explicitly maximizes the interdistributional distances during the optimization procedure. As a result, latent topics that have larger distances in their proportion distributions are favored. In addition, regularized by the first variational term, the optimized latent topics are also pertained to the Generative Process 2 and aimed to obtain the best representation of the documents.

Second, we maximize the group disparity in topic proportions via a classifier-free manner. Unlike the previous supervised topic models (Mcauliffe and Blei (2007), Card, Tan and Smith (2017)), which rely on a parameterized classifier to distinguish group identities, we evaluate groupwise topic proportion similarity by calculating the closed-form distance metrics directly from their distributions. At the groupwise distribution level, we optimize the variational free-parameters to increase the groupwise disparity. As a result, this approach not only relieves the need to posit and estimate extra parameters for the classifier but also protects the latent topics from any semantic distortion resulting from potential projections applied by the classifier. Moreover, by leveraging the counterfactual topic proportions, we no longer require separating topic proportion distribution for each

group and, therefore, can increase the sample efficiency and reduce the label imbalance level.

To conclude, we propose the HCF-DTM, which extends the VAE-NTM to a longitudinal setting and provides an efficient variational objective by segmenting the long dependent sequence into adjacent stages. To preserve the semantic meaning of the topics, HCF-DTM finds a number of time-consistent latent topics and maximizes groupwise heterogeneity via interdistributional distances. In the next section, we present the implementation details of the proposed method.

4. Implementation and algorithm. The optimization procedure focuses on two sets of parameters of interest: the generative parameters Ω and the variational parameters Ψ. Precisely, Ω specifies the Generative Process 2, including the time-invariant topics β and parameters $\phi_{1:T}$ of the mean trend functions $f_{1:T}$, whereas Ψ parameterizes the mean functions $g^{\mu}_{1:T}$ and variance functions $g^{\sigma}_{1:T}$ of the variational distribution Q. Our goal is to find the optimal set of Ω and Ψ, which maximize the value function presented in equation (7).

The implementation detail is summarized in Algorithm 1. It starts with learning the variational mean and variance, μ^q and σ^q, and then estimates the proposed value function which can be decomposed into three major components: the KL divergence between the variational and generative priors (μ^0 and σ^0), the log-likelihood of the documents generated from the latent topics, and the distributional distances between the counterfactual topic proportions. For demonstration purpose we consider a two-group setting and adopt the MI as the distance metrics. Calculating the log-likelihood and the groupwise MI is straightforward and can be conducted at each time stage. However, obtaining the longitudinal KL divergence is more challenging due to the temporal dependencies present in the follow-up stages. Though equation (6) breaks down the long temporal dependency to adjacent stages, it is still required to compute an expectation of KL divergence over the intractable space of all latent topic proportions at the previous stage, as shown in $\mathbb{E}_{\theta_{t-1}\sim q_{\psi_{t-1}}}(\mathbb{KL}[q_{\psi_t}(\theta_t|\boldsymbol{w}_t, X_t, Y)||p(\theta_t|\theta_{t-1}, X_t, Y)])$.

To address this challenge, we apply the reparametrization schemes M times to have a random sample of θ_{t-1} with size M drawn from the variational distribution. The sampled θ_{t-1} are later used to generate the prior distribution of θ_t based on the mean function f_t, which enables us to obtain an empirical estimate of the expected KL divergence at time t.

Algorithm 1 Heterogeneous Classifier-Free Dynamic Topic Model

1. **Initialize** generative parameters $\Omega = \{\beta, \phi_{1:T}\}$ and variational parameters $\Psi = \{\psi^{\mu}_{1:T}, \psi^{\sigma}_{1:T}\}$, variational mean functions $g^{\mu}_{1:T}$ and variance functions $g^{\sigma}_{1:T}$, topic proportion mean prior μ^0_1 and variance prior σ^0_1, KL sample number M, learning rate λ, maximum iterations $T_{\max}$, and a stopping error criterion ϵ_s.
2. **Input** all observed documents, metadata, and group $\{W_{i,1:T}, X_{i,1:T}, Y_i\}^N_{i=1}$.
3. **For** $k \leftarrow 1$ to $T_{\max}$, **do**
4. Compute variational $\mu^q_t = g^{\mu}_t(W_t, X_t, Y)$ and std.err $\sigma^q_t = g^{\sigma}_t(W_t, X_t, Y)$.
5. **For** $j \leftarrow 1$ to M, **do**
6. Sample Gaussian errors $\epsilon_{j,t} \sim N(0, 1)$ and reparametrize $\eta^q_{t,j} = \mu^q_t + \epsilon_{j,t} \cdot \sigma^q_t$.
7. Compute prior mean $\mu^0_{1,j} = f_1(\mu^0_1, X_1, Y)$, and $\mu^0_{t,j} = f_t(\eta^q_{t-1,j}, X_t, Y)$.
8. Compute counterfactual $\tilde{\mu}^q_t = g^{\mu}_t(W_t, X_t, 1-Y)$ and $\tilde{\sigma}^q_t = g^{\sigma}_t(W_t, X_t, 1-Y)$.
9. Compute gradient ∇ of equation (8) w.r.t. Ω and Ψ.
10. Update $(\Omega, \Psi)^k \leftarrow (\Omega, \Psi)^{k-1} - \lambda \cdot \nabla$.
11. Stop if $|\mathcal{L}^k - \mathcal{L}^{k-1}| \le \epsilon$.
12. **Return** estimated topics β and variational functions $\{g^{\mu}_t\}^T_{t=1}$ and $\{g^{\sigma}_t\}^T_{t=1}$.

Together with two other loss terms, the final objective function can be derived as follows:

$$
\begin{aligned}
&(\Omega^*, \Psi^*) \\
&= \underset{\Omega,\Psi}{\operatorname{argmin}} \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{M} \underbrace{\left\{ \log\left(\frac{\sigma^0}{\sigma^q_{i,t}} \right) + \frac{(\sigma^q_{i,t})^2 + (\mu^q_{i,t} - \mu^0_{i,j,t})^2)}{2 \cdot (\sigma^0)^2} - \frac{1}{2} \right\}}_{\text{Gaussian KL divergence term}} \\
&\quad \underbrace{- W_{i,t} \cdot \log\{\sigma(\eta^q_{i,t,j}) \cdot \sigma(\beta)\}}_{\text{Multinomial likelihood term}} \underbrace{- \frac{1}{2} \left\{ \log\left(\frac{\sigma^q_{i,t} + \tilde{\sigma}^q_{i,t}}{4 \cdot \sigma^q_{i,t} \cdot \tilde{\sigma}^q_{i,t}} \right) + \frac{(\mu^q_{i,t} - \tilde{\mu}^q_{i,t})^2}{\sigma^q_{i,t} + \tilde{\sigma}^q_{i,t}} + \frac{1}{2} \right\}}_{\text{Mutual Information term}},
\end{aligned} \tag{8}
$$

where $\epsilon_{t,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\eta^q_{i,t,j} = \mu^q_{i,t} + \epsilon_{t,j} \cdot \sigma^q_{i,t}$ is the unnormalized variational topic proportion sample and $\mu^0_{i,j,t} = f_t(\eta^q_{i,t-1,j}, X_{i,t}, Y_i)$ is the resulting prior proportion mean. Parameters of interest are updated jointly via stochastic gradient descent (Robbins and Monro (1951)). For further reference we present a graphical illustration of our model architecture and leave more detailed optimization choices, such as hyperparameter tuning, in the Supplementary Material (Ye et al. (2024)).

5. Simulation. In this section we present extensive simulation studies to illustrate the longitudinal interpretability and groupwise separation ability of our proposed method. To compare, we consider several BPTM-methods: LDA (Blei, Ng and Jordan (2003)), multistage dynamic LDA (mdLDA) (Blei and Lafferty (2006b)), and supervised LDA (sLDA) (Mcauliffe and Blei (2007)) as well as NTM-based approaches: prodLDA (Srivastava and Sutton (2017)) and SCHOLAR (Card, Tan and Smith (2017)). Note that sLDA and SCHOLAR are the two methods augmenting the document metadata, and mdLDA is the only multistage topic model which incorporates the longitudinal dependency of the documents. The main objective of the simulation is to investigate whether the topic models can recover the underlying topic distributions and identify subjects' group memberships under various settings, such as the number of time stages and topics.

The detailed simulation setting is described as follows. First, we consider a cohort of N subjects over T number of stages. At the baseline stage ($t = 1$), we assign 20 random features $\{X_{i1p}\}_{p=1}^{20}$ generated from a standard normal distribution $N(0,1)$ to each subject i. During all follow-up stages ($2 \le t \le T$), we let those features be correlated over time according to $X_{i,t,p} = X_{i,t-1,p} + \epsilon$, where $\epsilon \overset{\text{iid}}{\sim} N(0,1)$. The subjects are randomly assigned to one of the two groups $Y \in \{-1, 1\}$ with equal probabilities. After the subjects' metadata $(\{X_{i,t,.}\}_{t=1}^{T}, Y_i)_{i=1}^{N}$ are generated, we define the document-level generative priors, that is, topics β and proportions Θ.

Under each simulation setting, we assume the existence of K number underlying topics $\beta_{V \times K}$, where each topic $\beta_{1:V,k}$ describes a distribution over V number of words. For ease of denotations, the words are represented as numerical integer values ranging from 1 to V. To maximize the differences among topics, we let $\beta_{v,k} \sim \text{logit-Normal}(\mu_k, 1)$, where $\mu_k = \lfloor k \cdot \frac{V}{K} \rfloor$, and keep them constant across the time stages. Then we design functions $\{f_{t,1}, \ldots, f_{t,K}\}_{t=1}^{T}$ to generate time-varying topic proportions based on the document-level metadata. Specifically, the topics proportions at time t for subject i can be represented as $(\theta_{t,i,1}, \ldots, \theta_{t,i,K}) = \sigma(f_{t,1}(X_{i,t,.}, Y_i, \theta_{t-1,i,1}), \ldots, f_{t,K}(X_{i,t,.}, Y_i, \theta_{t-1,i,K}))$, where each function $f_{t,k}$ is parameterized into the following three components:

$$
f_{t,k} = \underbrace{\gamma^m_{t,k} \cdot X_{i,t,.}}_{\text{covariates main effect}} + \underbrace{\gamma^\theta_t \cdot \theta_{t-1,i,k}}_{\text{dependency from previous } \theta} + \underbrace{Y_i \cdot \gamma^g_{t,k} \cdot X_{i,t,.}}_{\text{group membership effect}}. \tag{9}
$$

The regression coefficients $\gamma_{t,k} \overset{\text{iid}}{\sim} N(0,1)$ are shared among all subjects at time t. In addition, equation (9) lists a linear relationship between the document metadata and the topic proportions. To increase the prior function complexity, we also impose nonlinearity by transforming metadata based on a specified function basis $\{X, X^2, X^3, \arctan X, \text{sign}(X)\}$. Finally, we generate documents $d_{i,t} \overset{\text{iid}}{\sim} \text{Mult}(\text{cnt}_{i,t}, \theta_{t,i,1:K} \cdot \beta^{\mathsf{T}}_{1:V,1:K})$ with number of words $\text{cnt}_{i,t} \overset{\text{iid}}{\sim} \text{Unif}(\{50, .., 150\})$.

We train the topic models on 80% of the generated documents and make inferences on a hold-out 1000 document set. Each topic model provides a posterior of the topic distributions $\hat{\beta}$ and proportions $\hat{\theta}$ and can be evaluated based on the following three criteria: topics recovery rate, dominant topics identification, and groupwise topic separation capability. All evaluation metrics are reported after 50 times of repeated experiments under each parameter setting. For demonstration purpose, we set $N = 1000$, $V = 200$, $T \in \{3, 5, 8\}$, $K \in \{3, 5, 8\}$, and generate topic proportions under nonlinear function setting. Additional parameter specifications and ablation studies are provided in Supplementary Material Section 3 (Ye et al. (2024)).

5.1. *Topics recovery rate.* To assess the model topic recovery rate, we compute the averaged empirical KL-divergence between the model estimated topic distributions $\hat{\beta}$ and ground-truth simulated topics β, that is,

$$\widehat{\mathbb{KL}}(\hat{\boldsymbol{\beta}}_{1:T} || \beta) = \frac{1}{K \cdot T} \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{v=1}^{V} \hat{\beta}_{t,v,k} \cdot \log(\beta_{v,k} / \hat{\beta}_{t,v,k}), \tag{10}$$

where a lower KL divergence indicates a smaller distributionalwise discrepancy and is more desired. However, due to the nature of unsupervised learning, the correspondence between the predicted and ground-true topics is undetermined. To incorporate this, we find the best permutation order $\mathcal{O}^*$ of the predicted topics, which reaches the minimum KL divergence metric, that is, $\mathcal{O}_t^* = \text{argmin}_{\mathcal{O}_t \in \text{perm}(1,..,K)} \widehat{\mathbb{KL}}(\hat{\beta}_{t,1:V,\mathcal{O}_t} || \beta_{1:V,1:K})$, as the recovered correspondence at each time stage t. Following the above procedure, we summarize the obtained KL metrics under each simulation setting in Table 1.

As shown in Table 1, HCF-DTM outperforms all other competing methods with respect to the topic distribution KL divergence. In particular, under a fixed number of topics, HCF-DTM obtains a better topics recovery rate, and the improvement margin enlarges compared to

TABLE 1

Empirical KL divergence between the estimated and actual topic word distributions when the generative prior function is nonlinear. Standard errors are summarized in the parentheses next to the estimated means. The improvement rate compares HCF-DTM against the best performer of the competing methods

K	T	LDA	sLDA	prodLDA	SCHOLAR	mdLDA	HCF-DTM	Imp-rate
3	3	15.326 (2.363)	19.319 (0.153)	22.066 (0.030)	15.663 (0.783)	9.633 (5.594)	**3.990** (0.429)	**58.580**%
	5	15.243 (2.278)	19.274 (0.207)	22.071 (0.024)	15.298 (0.528)	11.007 (5.933)	**3.805** (0.390)	**65.431**%
	8	15.334 (2.327)	19.293 (0.152)	22.072 (0.015)	15.022 (0.566)	14.438 (2.683)	**3.532** (0.395)	**75.537**%
5	3	18.976 (1.811)	23.154 (0.825)	27.007 (0.016)	21.843 (0.491)	13.042 (3.294)	**11.772** (1.454)	**9.738**%
	5	19.051 (1.864)	23.312 (0.711)	27.002 (0.015)	21.662 (0.615)	16.773 (3.001)	**9.017** (2.460)	**46.241**%
	8	19.037 (1.868)	23.066 (0.695)	27.003 (0.013)	21.548 (0.596)	18.708 (2.077)	**8.740** (2.877)	**53.282**%
8	3	23.384 (1.117)	27.544 (1.077)	31.004 (0.012)	27.097 (0.252)	23.737 (0.606)	**17.945** (1.629)	**23.259**%
	5	23.452 (1.051)	27.577 (0.578)	30.999 (0.010)	26.971 (0.239)	24.940 (0.670)	**15.280** (1.939)	**34.846**%
	8	23.592 (1.051)	27.815 (0.460)	31.001 (0.009)	26.763 (0.287)	26.491 (0.543)	**13.068** (2.041)	**44.608**%

the best-performing competing method (mdLDA) when the number of time stages increases. This illustrates the advantage of HCF-DTM which assumes a time-constant topic distribution by design. Such an assumption minimizes the number of inference parameters and eases the optimization procedure. Notably, though the single-stage topic model could adaptively search the optimal topics at each time stage, our time-consistent assumption, in alignment with the generative process, allows HCF-DTM to fully utilize documents across all time stages and thus makes it powerful to recover underlying consistent topics if there exists any.

5.2. *Dominant topic identification.* In this subsection we test model inference performance on topic proportions and evaluations according to the dominant topic alignment accuracy, that is,

$$\widehat{\mathrm{ACC}}_{\text{dom-topic}} = \frac{1}{T \cdot N} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{I}\{\operatorname{argmax}(\hat{\theta}_{t,i,1:K}) = \operatorname{argmax}(\theta_{t,i,1:K})\}. \tag{11}$$

A topic is considered dominant when its probabilistic proportion is the largest among all other topics for each document and, therefore, accurately identifying the dominant topics is essential, as the number of dominant topics at each time stage could be utilized to track the topic's popularity. After updating the order of the estimated topics with the sequential correspondence from previous simulations, we provide simulation results for $N = 1000$ and $K = 8$ in Figure 3.

Based on Figure 3, we notice that the proposed HCF-DTM method obtains the highest averaged dominant topic identification accuracy and, meanwhile, achieves the smallest standard errors, compared to all the rest competing methods. This result is expected due to the following three reasons. First, compared to the single-stage methods, the proposed HCF-DTM incorporates the dynamic document metadata, which utilizes the topic proportion generative function with the subject-level longitudinal information and is able to control potential confounders. Second, instead of letting the topic proportions share a common corpus-level hyperparameter as mdLDA, our method directly establishes temporal dependency on topic proportions from previous stages so that the information on dominant topics can be propagated into the future stages. Lastly, due to the mutual dependency between the topics and corresponding proportions, the topic proportions generated by HCF-DTM can be enhanced from the previously best-performing estimated topic distributions.

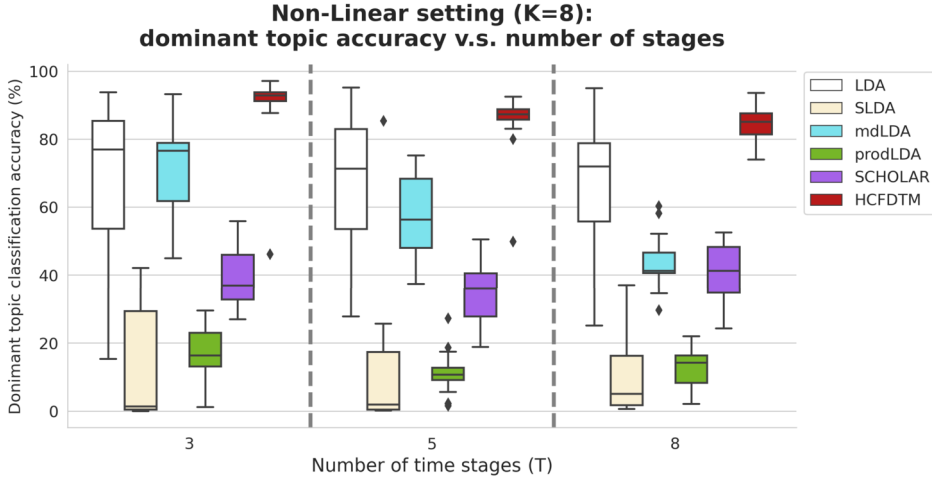

FIG. 3. *Boxplots of the dominant topic accuracy from estimated topic proportions vs. the number of time stages when* $N = 1000$, $K = 8$, *where the generative prior function is nonlinear. The left-to-right order of boxplot methods matches the top-to-bottom order of the legend.*

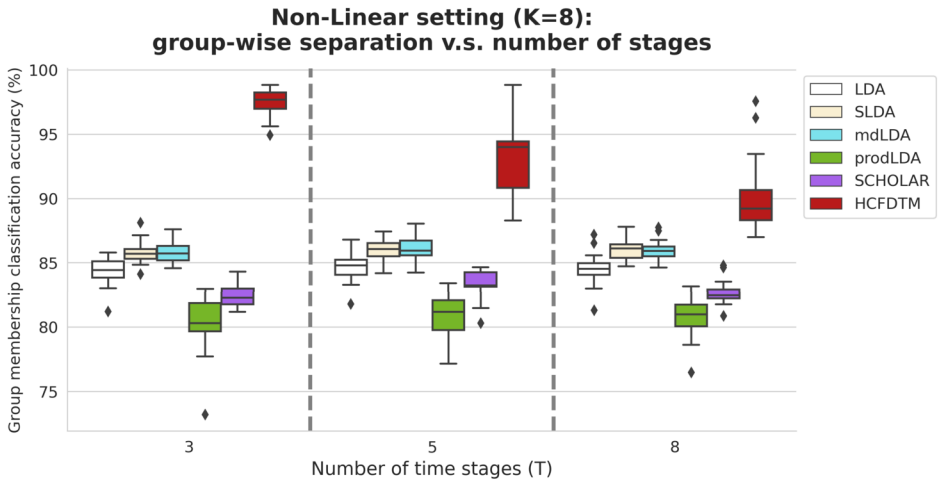

FIG. 4. *Boxplots of the group-membership accuracy from estimated topic proportions vs. the number of time stages when* $N = 1000$, $K = 8$, *and the generative prior function is nonlinear. The left-to-right order of boxplot methods matches the top-to-bottom order of the legend.*

5.3. *Groupwise topic separation.* In this simulation we examine the model capability of distinguishing topic proportions based on the subjects' group membership. In terms of evaluation metrics, we fit a simple logistic regression on the topic proportion posterior to predict subjects' group membership, that is, $Y_i \sim \text{Logistic}(\hat{\theta}_{t,i,1:K})$ and use the resulting classification accuracy to evaluate the modeling of groupwise topic separation capability. As the classification task is directly applied to the obtained topic proportions, a higher accuracy score indicates more groupwise information having been incorporated within the proportions and thus more desired. According to this criteria, we show the simulation results in Figure 4.

From Figure 4 we observe that, while the HCF-DTM model performance drops as the number of time stages increases, HCF-DTM still outperforms the rest of competing methods under each simulation setting. In particular, when there are three number of stages ($T = 3$), HCF-DTM improves the group-membership accuracy most substantially due to the small number of inference tasks. More importantly, according to the simulation results, we conclude that the proposed classifier-free approach to maximize the topic proportion differences among different groups is more efficient in embedding groupwise heterogeneity to the topic proportions compared to the additional classifiers imposed by the SCHOLAR and sLDA.

6. Real data analysis. In this section we apply the proposed topic model HCF-DTM to the clinical notes provided by the pediatric mental health department of the Children's Health of Orange County (CHOC). Over a five-year period from 01/10/2019 to 06/01/2023, CHOC has collected a total number of 25,957 notes from 2564 inpatient children with an average age of 14.48 years. These notes contain detailed contextual information about a patient, such as hospitalization reasons, initial nursing assessment, and conversations/interactions with the clinicians. The goal of this study is to understand the progression of children's mental health from the text data during COVID-19. In particular, we investigate any disparities in the SGM and non-SGM children's mental responses to stress factors associated with the pandemic.

In our analysis we first consolidate the time frame of interest by selecting three important events based on the timeline of COVID-19 (Centers for Disease Control and Prevention (2023)), which are 03/15/2020 when the states began to implement shutdown, 05/13/2021 when the vaccines were released and became available to the 12+ teenagers, and 08/01/2021 when the school reopen policy was announced. These three events segment the collection of

notes into four distinct time periods and are able to formulate the task into a four-stage longitudinal topic modeling setting. Additionally, as the sexual and gender identity of a patient is undisclosed, we search glossary terms, such as "pronouns" and "transgender" provided by the national LGBTQIA+ health education center, from the notes. This procedure reveals more than one-third (36%) of the children in our dataset potentially belonging to SGM. Remarkably, this number is much larger compared to the 2.2%–4.0% survey estimates of SGM proportion in the United States (Gates (2014)).

Within each specified time period, we preprocess the clinical notes according to standard procedure, including stemming, lemmatization, and removal of stop words. However, there still remain significant challenges. Due to the unstructured and distinctive characteristics of clinical notes, we encounter frequent spelling errors and extensive usage of medical terminology abbreviations, which largely increase the vocabulary size. To maintain a more manageable bag of words, we select terms which can be found in more than one-third of the notes. This filtering process results in a set of 273 instead of 37,890 number of unique words. Together with the tabular patients' demographic information and vital measurements, we apply the proposed HCF-DTM and competing methods to extract consistent themes from the notes to investigate inpatient children's mental status over the four stages of the pandemic.

In our application all methods are evaluated under two criteria: the accuracy in predicting children's sexual and gender identity based on the inferred topic proportions and the UCI coherence score (Newman et al. (2010)) for the extracted topics. Specifically, the first measures the extent of groupwise heterogeneity captured by the topic proportions, and the later score quantifies the semantic interpretability of the topics. To comprehensively analyze model performance, we randomly select 80% of the notes as a training set and repeat the process 20 times to obtain a Monte-Carlo sample of the model performance scores. The results are summarized in Table 2, where higher accuracy and coherence scores indicate better model performance.

According to Table 2, our HCF-DTM outperforms all other methods under each specification of the number of topics. Compared to the supervised models (sLDA and SCHOLAR) which rely on additional classifiers, our increased accuracy scores demonstrate that the introduced groupwise topic separation component of our model can be more effective in incorporating the sexual and gender identities of patients into the inferred topic proportions. Meanwhile, our augmentation of longitudinal covariate information further improves the obtained topics by capturing more enriched time-dependent heterogeneity among subjects. Furthermore, the proposed model's generative assumption of time-invariant topics, as opposed to the dynamically changing topics found by mdLDA, significantly enhances the interpretability of

TABLE 2

Model performance scores evaluated on the testing set of Monte-Carlo samples. Each test set consists of 1587 *notes from* 564 *patients. Standard errors are summarized in the parentheses next to the estimated means*

	K	LDA	sLDA	prodLDA	SCHOLAR	mdLDA	HCF-DTM
Accuracy	3	62.979%	65.119%	62.060%	65.693%	62.186%	**66.200**%
		(1.627)	(2.181)	(1.455)	(1.969)	(1.922)	(1.705)
	5	62.956%	65.642%	62.060%	65.456%	63.147%	**68.017**%
		(1.663)	(2.289)	(1.455)	(2.026)	(1.866)	(2.205)
	8	64.512%	66.592%	62.060%	65.759%	64.749%	**70.615**%
		(1.450)	(2.231)	(1.455)	(2.086)	(2.425)	(2.906)
Coherence	3	−0.162 (0.005)	−0.035 (0.038)	−0.344 (0.051)	0.096 (0.024)	−0.202 (0.023)	**0.152** (0.037)
	5	−0.091 (0.021)	0.099 (0.012)	−0.282 (0.044)	0.108 (0.110)	−0.219 (0.026)	**0.161** (0.058)
	8	−0.143 (0.007)	0.096 (0.020)	−0.246 (0.017)	0.164 (0.036)	−0.116 (0.027)	**0.181** (0.035)

FIG. 5. *Wordclouds of three topics extracted by HCF-DTM.*

the topics, indicated by the higher coherence scores. We summarize that our approach, combining document longitudinal metadata and group membership, efficiently addresses existing heterogeneity and thus attains the best performance in this data application.

Furthermore, with the highest coherence score achieved, we proceed to retrain HCF-DTM using the entire clinical note dataset with three specified topics for demonstration purposes. Detailed justification of this choice can be found in Supplementary Material Section 4.3 (Ye et al. (2024)). Figure 5 displays the wordclouds of the extracted topics that persist over time. Upon analyzing the word distributions, we assign the label "Interaction" to the first topic, due to the recurring words such as "parent, school," and "interaction" which describe the social support related to inpatient children. Similarly, we name the second and third topics "Positive" and "Negative," respectively, to reflect their positive or negative emotion and feelings. Finally, we identify the dominant topic with the highest topic proportion for each patient and time period. The dominant topic, which reflects a patient's primary mental status, illustrates the evolution of inpatient children's mental health over time. We first visualize the progression of the dominant topic proportions on a two-dimensional latent space extracted by t-SNE (Van der Maaten and Hinton (2008)) in Figure 6.

A direct observation of Figure 6 reveals that each topic is well separated and their relative positions on the two-dimensional latent space change dynamically across the four time periods. Notably, during the implementation of shutdown measures, the interdistances between three topics decrease while their variations increase, suggesting a significant impact on topic distributions from the shutdown event. Similarly, there is an increase in the "Negative" topic's variance in the latent space upon the reopening of schools. To further investigate how each topic progresses over the four COVID periods, we compute the percentage changes of the dominant topics and illustrate it in Figure 7.

Figure 7 shows there is an increase in the prevalence of "Negative" emotions and a decrease in the "Interaction" topics among both SGM and non-SGM children after the implementation of state shutdowns, whereas both trends reverse once schools began to reopen. In particular, compared to non-SGM children, SGM inpatient children exhibit more pronounced shifts, and their "Negative" emotions started to decline one stage earlier when news of vaccine availability was released. These disparities in topic trajectories suggest the existence of

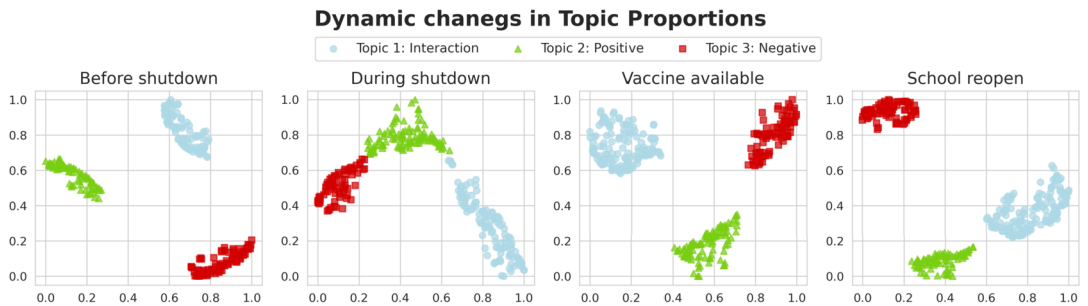

FIG. 6. *Dynamic changes in the topic proportions on a latent space extracted by the t-SNE. Each shape represents the corresponding dominant topic.*

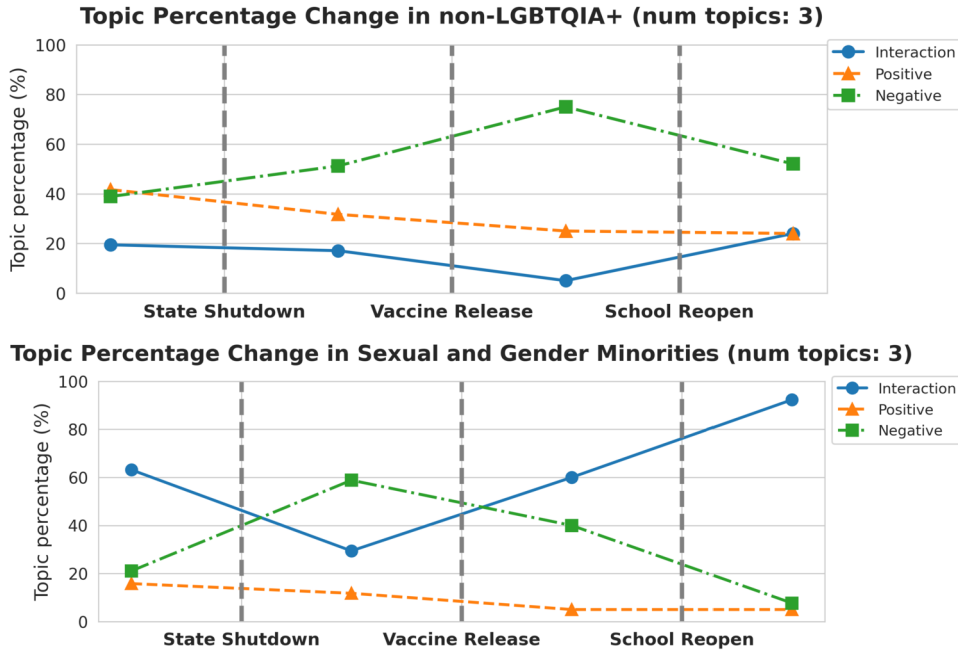

FIG. 7. *Percentage changes in the three topics fitted by HCFDTM within two patient groups according to sexual and gender identities.*

heterogeneity associated with children's sexual and gender identity. Importantly, our findings also align with recent research which shows that social isolation due to quarantine measures has led to elevated levels of anxiety but decreased vaccine hesitancy in the SGM community during COVID-19 (Pharr et al. (2022), Adzrago et al. (2023)).

To conclude, the proposed HCF-DTM method effectively extracts interpretable topics from the unstructured clinical notes. Our classifier-free groupwise separation approach reveals noticeable disparities in the progression of mental status between SGM and non-SGM children. This study effectively demonstrates the feasibility of using clinical notes to evaluate children's mental status and provides valuable insights for clinicians.

7. Discussion. In this paper we develop a heterogeneous dynamic topic model with an efficient variational inference procedure. The proposed model is designed to extract consistent topics in a multistage longitudinal setting. Specifically, our method maintains a set of time-invariant topics and incorporates document metadata into the topic proportions, where the first preserves the semantic meaning of each topic and the second captures the document temporal heterogeneity. In addition, when the documents can be categorized, we introduce a classifier-free topic learning approach that utilizes counterfactual topic distribution and inter-distributional distances to maximize topic heterogeneity across different document groups.

The proposed topic model is applied to clinical notes data from a large tertiary pediatric hospital in Southern California to evaluate inpatient children's mental health concerning their sexual and gender identities during the COVID-19 pandemic. We demonstrate three unique advantages of our method in this data application. First, without the need to navigate through a multitude of similar topics at each time stage, the extracted time-invariable topics readily represent children's mental status over time and enable us to quantify the children's mental health progression via the corresponding topic proportions. Second, the augmentation of documents metadata can efficiently incorporate the heterogeneity among inpatient children, such as their demographics features and evolving vital measurements. Lastly, our classifier-free groupwise heterogeneity maximization approach can effectively identify any disparity in

children's mental status related to their sexual and gender identities. Importantly, our model can be applied to other scenarios using topic modeling for longitudinal text data, particularly in the presence of heterogeneity.

Our real data analysis indicates that children tend to express more negative emotions during the state shutdowns and more positive when schools reopen. In particular, SGM children exhibit more pronounced reactions toward major COVID-19 events and greater sensitivity to vaccine-related news. This implies that increased social isolation due to enforced quarantine has a noticeable negative impact on children's mental health, especially among SGM children. As a result, engaging more social activities and support can be crucial for children's mental well-being, and our study may facilitate clinicians to understand the importance of rebuilding social connections and activities as a post-pandemic support system for children to recover from mental distress.

SUPPLEMENTARY MATERIAL

Supplementary Material to "Dynamic topic language model on heterogeneous children's mental health clinical notes" (DOI: 10.1214/24-AOAS1930SUPP; .pdf). The supplement contains variational inference proof, details of the optimization algorithm, additional simulations and ablation studies, and descriptions of real-world applications along with an examination of model assumptions.

REFERENCES

ADZRAGO, D., ORMISTON, C. K., SULLEY, S. and WILLIAMS, F. (2023). Associations between the self-reported likelihood of receiving the COVID-19 vaccine, likelihood of contracting COVID-19, discrimination, and anxiety/depression by sexual orientation. *Vaccines* **11** 582.

AFIFI, M. (2007). Gender differences in mental health. *Singapore Medical Journal* **48** 385.

BARRY, T. R. (2014). The midlife in the United States (MIDUS) series: A national longitudinal study of health and well-being. *Open Health Data* **2**.

BLEI, D. and LAFFERTY, J. (2006a). Correlated topic models. *Adv. Neural Inf. Process. Syst.* **18** 147.

BLEI, D. M. and LAFFERTY, J. D. (2006b). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* 113–120.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BOYD, A., GOLDING, J., MACLEOD, J., LAWLOR, D. A., FRASER, A., HENDERSON, J., MOLLOY, L., NESS, A., RING, S. et al. (2013). Cohort profile: The 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children. *Int. J. Epidemiol.* **42** 111–127.

CARD, D., TAN, C. and SMITH, N. A. (2017). Neural models for documents with metadata. Preprint. Available at arXiv:1705.09296.

CASALE, F. P., DALCA, A., SAGLIETTI, L., LISTGARTEN, J. and FUSI, N. (2018). Gaussian process prior variational autoencoders. *Adv. Neural Inf. Process. Syst.* **31**.

CIECHANOWSKI, K., JEMIELNIAK, D. and SILCZUK, A. (2023). Public interests in mental health topics in COVID-19: Evidence from Wikipedia searches. *Adv. Mental Health* 1–22.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2023). CDC Museum COVID-19 Timeline.

FORTUIN, V., BARANCHUK, D., RÄTSCH, G. and MANDT, S. (2020). Gp-vae: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics* 1651–1661. PMLR.

GATES, G. J. (2014). LGBT demographics: Comparisons among population-based surveys.

GULRAJANI, I., KUMAR, K., AHMED, F., TAIGA, A. A., VISIN, F., VAZQUEZ, D. and COURVILLE, A. (2016). Pixelvae: A latent variable model for natural images. Preprint. Available at arXiv:1611.05013.

GUPTA, P., CHAUDHARY, Y., BUETTNER, F. and SCHÜTZE, H. (2019). Document informed neural autoregressive topic models with distributional prior. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 6505–6512.

HU, X., WANG, R., ZHOU, D. and XIONG, Y. (2020). Neural topic modeling with cycle-consistent adversarial training. Preprint. Available at arXiv:2009.13971.

KARIM, S., CHOUKAS-BRADLEY, S., RADOVIC, A., ROBERTS, S. R., MAHEUX, A. J. and ESCOBAR-VIERA, C. G. (2022). Support over social media among socially isolated sexual and gender minority youth in rural US during the COVID-19 pandemic: Opportunities for intervention research. *Int. J. Environ. Res. Public Health* **19** 15611.

LAROCHELLE, H. and LAULY, S. (2012). A neural autoregressive topic model. *Adv. Neural Inf. Process. Syst.* **25**.
LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791. https://doi.org/10.1038/44565
LI, X., OUYANG, J. and ZHOU, X. (2015). Supervised topic models for multi-label classification. *Neurocomputing* **149** 811–819.
LI, Y., ZHU, R., QU, A., YE, H. and SUN, Z. (2021). Topic modeling on triage notes with semiorthogonal non-negative matrix factorization. *J. Amer. Statist. Assoc.* **116** 1609–1624. MR4353700 https://doi.org/10.1080/01621459.2020.1862667
LIN, T., HU, Z. and GUO, X. (2019). Sparsemax and relaxed Wasserstein for topic sparsity. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* 141–149.
MARSHAL, M. P., DIETZ, L. J., FRIEDMAN, M. S., STALL, R., SMITH, H. A., MCGINLEY, J., THOMA, B. C., MURRAY, P. J., D'AUGELLI, A. R. et al. (2011). Suicidality and depression disparities between sexual minority and heterosexual youth: A meta-analytic review. *J. Adolesc. Health* **49** 115–123. https://doi.org/10.1016/j.jadohealth.2011.02.005
MCAULIFFE, J. and BLEI, D. (2007). Supervised topic models. *Adv. Neural Inf. Process. Syst.* **20**.
MCGEOUGH, B. L. and STERZING, P. R. (2018). A systematic review of family victimization experiences among sexual minority youth. *J. Prim. Prev.* **39** 491–528. https://doi.org/10.1007/s10935-018-0523-x
MCGREGOR, K., WILLIAMS, C. R., BOTTA, A., MANDEL, F. and GENTILE, J. (2023). Providing essential gender-affirming telehealth services to transgender youth during COVID-19: A service review. *J. Telemed. Telecare* **29** 147–152. https://doi.org/10.1177/1357633X221095785
MIAO, Y., YU, L. and BLUNSOM, P. (2016). Neural variational inference for text processing. In *International Conference on Machine Learning* 1727–1736. PMLR.
NEWMAN, D., NOH, Y., TALLEY, E., KARIMI, S. and BALDWIN, T. (2010). Evaluating topic models for digital libraries. In *Proceedings of the* 10*th Annual Joint Conference on Digital Libraries* 215–224.
NATIONAL INSTITUTE OF MENTAL HEALTH (2021). Mental health topics. From https://www.nimh.nih.gov/health/topics.
PAATERO, P. and TAPPER, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5** 111–126.
PENNINX, B. W., BEEKMAN, A. T., SMIT, J. H., ZITMAN, F. G., NOLEN, W. A., SPINHOVEN, P., CUIJPERS, P., DE JONG, P. J., VAN MARWIJK, H. W. et al. (2008). The Netherlands Study of Depression and Anxiety (NESDA): Rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17** 121–140.
PHARR, J. R., TERRY, E., WADE, A., HABOUSH-DELOYE, A., MARQUEZ, E., HEALTH, N. M. and COALITION, E. (2022). Impact of COVID-19 on sexual and gender minority communities: Focus group discussions. *Int. J. Environ. Res. Public Health* **20** 50.
PLÖDERL, M. and TREMBLAY, P. (2015). Mental health of sexual minorities. A systematic review. *Int. Rev. Psychiatry* **27** 367–385. https://doi.org/10.3109/09540261.2015.1083949
RAMCHANDRAN, S., TIKHONOV, G., KUJANPÄÄ, K., KOSKINEN, M. and LÄHDESMÄKI, H. (2021). Longitudinal variational autoencoder. In *International Conference on Artificial Intelligence and Statistics* 3898–3906. PMLR.
RAVENS-SIEBERER, U., KAMAN, A., ERHART, M., DEVINE, J., SCHLACK, R. and OTTO, C. (2022). Impact of the COVID-19 pandemic on quality of life and mental health in children and adolescents in Germany. *Eur. Child Adolesc. Psychiatry* **31** 879–889.
ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586
ROBERTS, M. E., STEWART, B. M., TINGLEY, D., LUCAS, C., LEDER-LUIS, J., GADARIAN, S. K., ALBERTSON, B. and RAND, D. G. (2014). Structural topic models for open-ended survey responses. *Amer. J. Polit. Sci.* **58** 1064–1082.
ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512.
RONALD, W., CAROL, D., ELSIE, J., LELA, R., SATVINDER, D. and TARA, W. (2010). Evolving definitions of mental illness and wellness. *Prev. Chronic Dis.* **7** 2.
ROSENFIELD, S. and MOUZON, D. (2013). Gender and mental health. *Handbook of the Sociology of Mental Health* 277–296.
RUSSELL, S. T. and FISH, J. N. (2016). Mental health in lesbian, gay, bisexual, and transgender (LGBT) youth. *Annu. Rev. Clin. Psychol.* **12** 465–487. https://doi.org/10.1146/annurev-clinpsy-021815-093153
SALERNO, J. P., DEVADAS, J., PEASE, M., NKETIA, B. and FISH, J. N. (2020). Sexual and gender minority stress amid the COVID-19 pandemic: Implications for LGBTQ young persons' mental health and well-being. *Public Health Rep.* **135** 721–727. https://doi.org/10.1177/0033354920954511
SCOTT, W. A. (1958). Research definitions of mental health and mental illness. *Psychol. Bull.* **55** 29.

SGARRO, A. (1981). Informational divergence and the dissimilarity of probability distributions. *Calcolo* **18** 293–302. MR0647828 https://doi.org/10.1007/BF02576360

SHARIFIAN-ATTAR, V., DE, S., JABBARI, S., LI, J., MOSS, H. and JOHNSON, J. (2022). Analysing longitudinal social science questionnaires: Topic modelling with BERT-based embeddings. In 2022 *IEEE International Conference on Big Data* (*Big Data*) 5558–5567. IEEE.

SIBSON, R. (1969/70). Information radius. *Z. Wahrsch. Verw. Gebiete* **14** 149–160. MR0258198 https://doi.org/10.1007/BF00537520

SRIDHAR, D., DAUMÉ III, H. and BLEI, D. (2022). Heterogeneous supervised topic models. *Trans. Assoc. Comput. Linguist.* **10** 732–745.

SRIVASTAVA, A. and SUTTON, C. (2017). Autoencoding variational inference for topic models. Preprint. Available at arXiv:1703.01488.

THOMA, B. C., REZEPPA, T. L., CHOUKAS-BRADLEY, S., SALK, R. H. and MARSHAL, M. P. (2021). Disparities in childhood abuse between transgender and cisgender adolescents. *Pediatrics* **148**. https://doi.org/10.1542/peds.2020-016907

THOMPSON, L. and MIMNO, D. (2020). Topic modeling with contextualized word representation clusters. Preprint. Available at arXiv:2010.12626.

VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**.

WANG, C., BLEI, D. and HECKERMAN, D. (2012). Continuous time dynamic topic models. Preprint. Available at arXiv:1206.3298.

WANG, R., ZHOU, D. and HE, Y. (2019). Atm: Adversarial-neural topic model. *Inf. Process. Manag.* **56** 102098.

WHAIBEH, E., VOGT, E. L. and MAHMOUD, H. (2022). Addressing the behavioral health needs of sexual and gender minorities during the COVID-19 pandemic: A review of the expanding role of digital health technologies. *Curr. Psychiatry Rep.* **24** 387–397. https://doi.org/10.1007/s11920-022-01352-1

WU, T., JIA, X., SHI, H., NIU, J., YIN, X., XIE, J. and WANG, X. (2021). Prevalence of mental health problems during the COVID-19 pandemic: A systematic review and meta-analysis. *J. Affective Disorders* **281** 91–98.

YE, H., MORENO, T., ALPERN, A., EHWERHEMUEPHA, L. and QU, A. (2024). Supplement to "Dynamic topic language model on heterogeneous children's mental health clinical notes." https://doi.org/10.1214/24-AOAS1930SUPP