Journal of Machine Learning Research 17 (2016) 1-38 Submitted 8/15; Revised 4/16; Published 7/16

Cross-Corpora Unsupervised Learning of Trajectories in Autism Spectrum Disorders

Huseyin Melih Elibol ELIBOL@G.HARVARD.EDU
Vincent Nguyen VINCENTNGUYEN@ALUMNI.HARVARD.EDU
Scott Linderman SLINDERMAN@SEAS.HARVARD.EDU
Matthew Johnson MATTJJ@CSAIL.MIT.EDU
Amna Hashmi AMNAHASHMI@ALUMNI.HARVARD.EDU
Finale Doshi-Velez FINALE@SEAS.HARVARD.EDU
Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138, USA

Editor: Benjamin M. Marlin, C. David Page, and Suchi Saria

Abstract

Patients with developmental disorders, such as autism spectrum disorder (ASD), present with symptoms that change with time even if the named diagnosis remains fixed. For example, language impairments may present as delayed speech in a toddler and difficulty reading in a school-age child. Characterizing these trajectories is important for early treatment. However, deriving these trajectories from observational sources is challenging: electronic health records only reflect observations of patients at irregular intervals and only record what factors are clinically relevant at the time of observation. Meanwhile, caretakers discuss daily developments and concerns on social media.

In this work, we present a fully unsupervised approach for learning disease trajectories from incomplete medical records and social media posts, including cases in which we have only a single observation of each patient. In particular, we use a dynamic topic model approach which embeds each disease trajectory as a path in $\mathbb{R}^D$. A Pólya-gamma augmentation scheme is used to efficiently perform inference as well as incorporate multiple data sources. We learn disease trajectories from the electronic health records of 13,435 patients with ASD and the forum posts of 13,743 caretakers of children with ASD, deriving interesting clinical insights as well as good predictions.

Keywords: Disease progression model, Dynamic topic model

1. Introduction

Psychiatric conditions that arise in childhood, generally termed developmental disorders, are increasingly common. The parent-reported rates of development disorders are now nearly 15%, which includes learning disabilities (affecting 7.66% of children) and attention deficit hyperactivity disorder (ADHD, 6.69% of children) (Boyle et al., 2011). CDC estimates for the prevalence of autism spectrum disorder (ASD) is now 1 in 68 children which is over 1% of the US population (Baio, 2014).

Characterizing these disorders is challenging because, unlike many adult disorders, the symptoms of developmental disorders are inextricably linked to the developmental processes

of childhood. For example, a language-related impairment may present as delayed speech in a toddler and difficulty reading in a school-age child. A neurological condition may manifest as convulsions at age three and intellectual disability at age seven. Characterizing the evolution of distinct disease courses is a critical step toward personalizing treatments; with developmental disorders the early identification of appropriate therapy can significantly increase a child's IQ and ability to communicate (Peters-Scheffer et al., 2011).

However, constructing these trajectories from data is challenging. Clinical studies tend to have the cleanest sources of data: patients are followed regularly, and measurements are consistently recorded. Unfortunately, most clinical studies involve small cohorts—under 200 individuals—which can make it difficult or impossible to distinguish heterogeneous disease courses from variance. In contrast, electronic health records (EHRs) and social media (SM) provide valuable windows to study populations of thousands of individuals. However, these less-structured sources are much more challenging to analyze due to several factors:

- *(Extremely) Partial trajectories.* EHRs are often confined to a single medical system; if a patient switches providers then their history will no longer be available. Similarly, patients and caregivers may be active on social media at some times and not others.
- *Irregular interactions.* Patients generally only visit clinics or post to social media when they have complaints; we do not observe data from patients between these times.
- *Partially structured, noisy, high-dimensional information.* The space of clinical symptoms is large, and with both clinician and caregiver-generated text, information may also be entered or described incorrectly. Clinicians and patients use very different vocabularies when describing the same symptoms.

To address these challenges, we develop an unsupervised approach that models each source—electronic health records and social media—with a cross-corpora dynamic topic model. Our model can be scientifically interpreted as positing that there are a few underlying disease processes that characterize the signs and symptoms that we observe in our patient population. Each disease is a process that evolves over time; we posit that each disease process k at each time t is associated with a distribution over possible signs and symptoms it may emit. The same disease process may be described differently in electronic health records and social media, and multiple diseases may be simultaneously present in a patient.

Specifically, we assume data in the form of (patient , time , sign) tuples. For some patients, we have may have data at multiple times; for other patients, we may only have data at one time. Similarly, some patients may have many signs, others just a few. Our approach derives distinct disease trajectories *without* linking individual identities between social media and electronic health records, and it can also derive disease trajectories in the limit of only a single note per patient. Thus, we do not have to restrict ourselves to patients with longitudinal data; we are able to incorporate all patient data that we have.

For inference in our model, we explore the use a Pólya-gamma augmentation scheme (Polson et al., 2013; Zhou et al., 2012b; Chen et al., 2013; Linderman et al., 2015) to easily adapt the model to have different correlation structures. We detail our approach in Sections 3 and 4, and review related work in Section 6. In Section 5, we apply our approach to a large data set of electronic health records from 13,435 individuals with ASD and 13,743

forum posts by 2,391 caretakers of children with ASD. To our knowledge, this is the first study to jointly model temporal patterns in electronic health record and social media data at this scale.

2. Background

Our technical approach uses Pólya-gamma augmentation to construct an efficient and easily extensible sampler for dynamic topic models and related models. In this section we briefly review topic models and Pólya-gamma augmentation.

2.1 Topic Models and Dynamic Topic Models

Latent Dirichlet Allocation (LDA) The latent Dirichlet allocation (LDA) topic model (Blei et al., 2003) is one of the most successful and widely used models in machine learning. Its basic aim is to decompose a corpus of natural language documents, like a collection of news articles or scientific papers, into an interpretable collection of topics as well as identify what topics are present in each document. For example, a corpus of scientific papers may contain topics like atomic physics, cosmology, and neural chemistry. For modeling purposes, each such topic is identified with a distribution over words: for example, the word "experiment" might have high probability in all three topics, while only the cosmology topic might have frequent occurrences of words like "star" and "galaxy." In this simplified view, to identify the topics present in a document, it is not necessary to model the details of language or even the order of the words in each document; instead, a document can be summarized by "bag of words:" a histogram counting the words that it contains.

The LDA topic model of Blei et al. (2003) posits that each document can be characterized by a distribution over the topics it contains, and each topic can be characterized by a distribution over the words associated with it. In symbols, each document d has a distribution over topics $\boldsymbol{\theta}_d$ ($d = 1, 2, \ldots, D$), and each topic $\boldsymbol{\beta}_k$ ($k = 1, 2, \ldots, K$) is a distribution over a vocabulary of V possible words. Given Dirichlet priors on the topics $\boldsymbol{\beta}$ and topic proportions $\boldsymbol{\theta}$ with parameters α_β and α_θ, the full generative model (also illustrated in figure 1a) is

$$\begin{aligned} \boldsymbol{\beta}_k &\sim \text{Dir}(\alpha_\beta), \\ \boldsymbol{\theta}_d &\sim \text{Dir}(\alpha_\theta), \\ z_{n,d} \,|\, \boldsymbol{\theta}_d &\sim \text{Cat}(\boldsymbol{\theta}_d), \\ w_{n,d} \,|\, z_{n,d}, \{\boldsymbol{\beta}_t\} &\sim \text{Cat}(\boldsymbol{\beta}_{z_{n,d}}). \end{aligned} \tag{1}$$

where $w_{n,d}$ is n^{th} word in document d, $z_{n,d}$ is the topic associated with the word $w_{n,d}$, $\text{Cat}(\boldsymbol{\pi})$ draws one sample from a vector of probabilities $\boldsymbol{\pi}$, and Dir is the Dirichlet distribution. The Dirichlet-multinomial conjugacy in the generative process makes it straight-forward to perform inference via a blocked Gibbs sampling scheme that, given a set of words $\{w_{n,d}\}$, can sample the latent topic-word distributions $\{bbeta_k\}$, the document-topic proportions $\{\boldsymbol{\theta}_d\}$, and the word-topic assignments $\{z_{n,d}\}$.

Dynamic Topic Model (DTM) Blei and Lafferty (2006b) expand upon LDA to model temporal evolution in the topics $\boldsymbol{\beta}$. Each multinomial topic distribution $\boldsymbol{\beta}_k$ is modeled

through its natural parameter $\boldsymbol{\psi}_k$; the mapping from $\boldsymbol{\psi}_k$ to $\boldsymbol{\beta}_k$ is a multi-class logistic function given by

$$\boldsymbol{\beta}_k(v) \equiv \boldsymbol{\beta}(\boldsymbol{\psi}_k(v)) \equiv \frac{\exp(\boldsymbol{\psi}_k(v))}{\sum_{v'} \exp(\boldsymbol{\psi}_k(v'))}. \tag{2}$$

where $\boldsymbol{\beta}_k(v)$ is the probability of word v in topic k. The natural parameters $\boldsymbol{\psi}_k$ are unconstrained—they can be positive or negative, and they do not need to sum to one.

Next, Blei and Lafferty (2006b) model the evolution of each topic $\boldsymbol{\beta}_k$ as a random walk on its natural parameters $\boldsymbol{\psi}_k$. Let $\boldsymbol{\psi}_{k,t}$ denote the values of the natural parameters $\boldsymbol{\psi}$ for topic k at time t. The DTM posits the following generative process on $\boldsymbol{\psi}$, also illustrated in figure 1b:

$$\begin{aligned}
\boldsymbol{\psi}_{k,t} | \boldsymbol{\psi}_{k,t-1} &\sim \mathcal{N}(\boldsymbol{\psi}_{k,t-1}, \sigma^2 I), \\
\boldsymbol{\theta}_d &\sim \mathrm{Dir}(\alpha_\theta), \\
z_{n,d} \,|\, \boldsymbol{\theta}_d &\sim \mathrm{Cat}(\boldsymbol{\theta}_d), \\
w_{n,d} \,|\, z_{n,d}, \{\boldsymbol{\psi}_{t,k}\} &\sim \mathrm{Cat}(\boldsymbol{\beta}(\boldsymbol{\psi}_{z_{n,d},t(d)})).
\end{aligned} \tag{3}$$

Here, $t(d)$ is the time associated with document d and $\boldsymbol{\beta}(\boldsymbol{\psi}_{k,t})$ is the transformation of $\boldsymbol{\psi}_{k,t}$ back to a multinomial using equation 2. We will use $\boldsymbol{\beta}_{k,t}$ as shorthand for $\boldsymbol{\beta}(\boldsymbol{\psi}_{k,t})$.

This DTM construction captures the temporal evolution of topics while retaining the interpretable structure of LDA. However, the DTM construction in equation 3 does not enjoy the conjugacy structure of the original LDA model in equation 1: the DTM replaces LDA's factored Dirichlet prior on the topics $\boldsymbol{\beta}_k$ with a Gaussian linear dynamical system (LDS) mapped through a multi-class logistic function. While inference in Gaussian linear dynamical systems coupled with linear Gaussian observations can be performed efficiently using message passing algorithms, the nonlinear mapping in equation 2 does not allow such algorithms to be applied directly.

2.2 Pólya-gamma Augmentation

Pólya-gamma augmentation is an auxiliary variable scheme that allows multinomial observations to appear as Gaussian likelihoods. This scheme has recently been used to develop Gibbs samplers and variational inference algorithms for Bernoulli, binomial, negative binomial, and multinomial regression models with logit link functions (Polson et al., 2013). Chen et al. (2013) use Pólya-gamma augmentation for multinomial models in the context of LDA, but in a way that only provides limited single-site inference updates. More recently, Linderman et al. (2015) extend the Pólya-gamma augmentation scheme for multinomial models in such a way that allows block updates and hence readily extends to dynamic topic models, in which entire state trajectories must be updated as a block for inference to be efficient. Here, we use the augmentation strategy of Linderman et al. (2015) to enable such block updating in our dynamic topic models.

The Pólya-gamma augmentation scheme is based on an integral identity derived from the Laplace transform of the Pólya-gamma distribution. Specifically, if $p(\omega \,|\, b, 0)$ is the density of the Pólya-gamma distribution $\mathrm{PG}(b, 0)$, then

$$\frac{(e^\psi)^a}{(1+e^\psi)^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2}\, p(\omega \,|\, b, 0)\, \mathrm{d}\omega, \tag{4}$$

where $\kappa = a - b/2$. The integral on the right-hand side is the Laplace transform of the Pólya-gamma density evaluated at $\psi^2/2$, and the left-hand side is a functional form that often appears in logistic likelihoods. Importantly, viewed as a function of ψ for fixed ω, the right-hand side is an unnormalized Gaussian density. Thus, the identity in equation 4 transforms a logistic likelihood to a Gaussian likelihood conditioned on an auxiliary variable, ω.

While we focus on Gibbs sampling inference here, the Pólya-gamma augmentation scheme also enables efficient mean-field variational inference (Linderman et al., 2015; Zhou et al., 2012b), including scalable stochastic variational inference (Hoffman et al., 2013; Linderman et al., 2015). These algorithms could be adapted to provide scalable inference for the dynamic topic model case that we study here.

Binomial Case For the binomial case, Polson et al. (2013) let $\psi_0 = 0$ and write $\psi_1 = \psi$. Let $x = (x_0, x_1)$ be the number of zeros and ones that have been observed. Then we can write the likelihood of the natural parameter ψ given the data x as

$$p(x \,|\, \psi) = \binom{x_0 + x_1}{x_1} \frac{(e^\psi)^{x_1}}{(1 + e^\psi)^{x_0}} = c(x) \frac{(e^\psi)^{a(x)}}{(1 + e^\psi)^{b(x)}}$$

Given a prior $p(\psi)$ on the natural parameter ψ, then the joint density of (ψ, x) can be written as

$$p(\psi, x) = p(\psi)\, c(x) \frac{(e^\psi)^{a(x)}}{(1 + e^\psi)^{b(x)}} = \int_0^\infty p(\psi)\, c(x)\, 2^{-b(x)} e^{\kappa(x)\psi} e^{-\omega\psi^2/2}\, p(\omega \,|\, b(x), 0)\, \mathrm{d}\omega. \quad (5)$$

The integrand of (5) defines a joint density on (ψ, x, ω). If we condition on the auxiliary variables ω, then the conditional density $p(\psi \,|\, x, \omega)$ on the natural Bernoulli parameter ψ is given by

$$p(\psi \,|\, x, \omega) \propto p(\psi) e^{\kappa(x)\psi} e^{-\omega\psi^2/2} \quad (6)$$

which is Gaussian if $p(\psi)$ is Gaussian. By the exponential tilting property of the Pólya-gamma distribution, we have $\omega \,|\, \psi, x \sim \mathrm{PG}(b(x), \psi)$. Efficient samplers exist for Pólya-gamma distributed variables (Windle et al., 2014), and thus we can alternate between sampling $\omega \,|\, \psi, x$ from a Pólya-gamma distribution and sampling $\psi \,|\, \omega, x$ from a Gaussian distribution.

Multinomial Case For the multinomial case, Linderman et al. (2015) rewrite the K-dimensional multinomial likelihood recursively in terms of $K - 1$ binomial densities using the following stick-breaking representation. Let $\boldsymbol{\beta}$ be a vector describing the probability of each outcome $1 \ldots K$. Then we can define $\widetilde{\boldsymbol{\beta}}_k$ to be probability of choosing option k given that we have not selected any option $j < k$:

$$\widetilde{\boldsymbol{\beta}}_k = \frac{\boldsymbol{\beta}_k}{1 - \sum_{j<k} \boldsymbol{\beta}_j} \quad (7)$$

Writing the probabilities $\boldsymbol{\beta}$ in this way allows us to write the multinomial density as a product of binomial densities:

$$\text{Mult}(\boldsymbol{x} \mid N, \boldsymbol{\beta}) = \prod_{k=1}^{K-1} \text{Bin}(x_k \mid N_k, \widetilde{\boldsymbol{\beta}}_k), \tag{8}$$

$$N_k = N - \sum_{j<k} x_j, \quad k = 1, 2, \ldots, K, \tag{9}$$

where we can interpret N_k as the number of observations remaining after the observations where $j = 1, 2, \ldots, k$ have been removed. Substituting $\widetilde{\boldsymbol{\beta}}_k = \sigma(\boldsymbol{\psi}_k)$, we can write the multinomial likelihood as

$$\begin{aligned}\text{Mult}(\boldsymbol{x} \mid N, \boldsymbol{\psi}) = \prod_{k=1}^{K-1} \text{Bin}(x_k \mid N_k, \sigma(\boldsymbol{\psi}_k)) &= \prod_{k=1}^{K-1} \binom{N_k}{x_k} \sigma(\boldsymbol{\psi}_k)^{x_k} (1 - \sigma(\boldsymbol{\psi}_k))^{N_k - x_k} \\ &= \prod_{k=1}^{K-1} \binom{N_k}{x_k} \frac{(e^{\boldsymbol{\psi}_k})^{x_k}}{(1 + e^{\boldsymbol{\psi}_k})^{N_k}}.\end{aligned}$$

Linderman et al. (2015) next let $\boldsymbol{a}_k(\boldsymbol{x}) = \boldsymbol{x}_k$ and $\boldsymbol{b}_k(\boldsymbol{x}) = N_k$ for each $k = 1, 2, \ldots, K-1$ and introduce Pólya-gamma auxiliary variables $\boldsymbol{\omega}_k$ corresponding to each coordinate $\boldsymbol{\psi}_k$. Then the probability of the data $\boldsymbol{x}$ and the auxiliary variables $\boldsymbol{\omega}$ given the natural parameters $\boldsymbol{\psi}$ has a diagonal Gaussian likelihood:

$$p(\boldsymbol{x}, \boldsymbol{\omega} \mid \boldsymbol{\psi}) \propto \prod_{k=1}^{K-1} e^{(x_k - N_k/2)\psi_k - \omega_k \psi_k^2/2} \propto \mathcal{N}\left(\boldsymbol{\psi} \,\middle|\, \boldsymbol{\Omega}^{-1}\kappa(\boldsymbol{x}),\, \boldsymbol{\Omega}^{-1}\right),$$

where $\boldsymbol{\Omega} \equiv \text{diag}(\boldsymbol{\omega})$ and $\kappa(\boldsymbol{x}) \equiv \boldsymbol{x} - N(\boldsymbol{x})/2$. Thus, if we begin with a Gaussian prior $p(\boldsymbol{\psi})$ on the stick-breaking parameters $\boldsymbol{\psi}$, then the posterior will remain Gaussian.

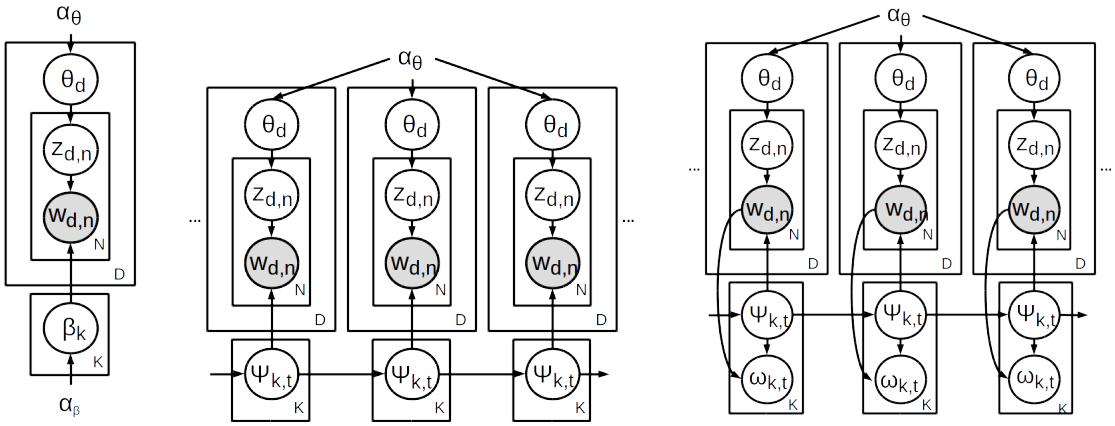
Finally, given the parameters $\boldsymbol{\psi}$, we can recover the parameters $\boldsymbol{\beta}$ through the stick-breaking construction:

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_j &= \sigma(\boldsymbol{\psi}_j) \\ \boldsymbol{\beta}_k &= \widetilde{\boldsymbol{\beta}}_k \prod_{j<k} (1 - \widetilde{\boldsymbol{\beta}}_j)\end{aligned} \tag{10}$$

We denote this recovery process in equation 10 by the function $\boldsymbol{\beta} \equiv \pi_{\mathsf{SB}}(\boldsymbol{\psi})$.

3. Model: Stick-breaking Construction for Dynamic Topic Models

The Pólya-gamma augmentation scheme allows us to take a Gaussian graphical model in which efficient inference is well-developed and apply it to models with multinomial likelihoods. However, we must first convert the dynamic topic model from Section 2.1 into the appropriate stick-breaking form. In this section we describe this stick-breaking construction and a natural cross-corpora extension; for completeness we also include the parts of the dynamic topic model that remain unchanged.



(a) LDA (b) DTM (c) Stick-Breaking DTM

Figure 1: Graphical Models of latent Dirichlet allocation, the dynamic topic model, and our stick-breaking dynamic topic model. The natural parameters $\boldsymbol{\psi}$ are converted to multinomials $\boldsymbol{\beta}$ through the stick-breaking process in equation 10

3.1 Document-Specific parameters $\{\boldsymbol{\theta}_d\}$ and $\{z_{n,d}\}$

As in the standard LDA approach, we continue to model the proportion of each topic in each document $\boldsymbol{\theta}_d$ as being drawn independently from Dirichlet distributions with parameters α_θ, and the topic $z_{n,d}$ for each word $w_{n,d}$ drawn from $\boldsymbol{\theta}_d$:

$$\begin{aligned} \boldsymbol{\theta}_d &\sim \text{Dir}(\alpha_\theta), \\ z_{n,d} \,|\, \boldsymbol{\theta}_d &\sim \text{Cat}(\boldsymbol{\theta}_d). \end{aligned}$$

3.2 Topic Parameters $\{\boldsymbol{\beta}_k\}$

Static Stick-Breaking LDA Model In standard LDA, the likelihood associated with each topic $\boldsymbol{\beta}_k$ depends on the words assigned to that topic:

$$p(\{\boldsymbol{w}_d\}_{d=1}^D \,|\, , \{\boldsymbol{z}_d\}_{d=1}^D, \{\boldsymbol{\beta}_k\}_{k=1}^K) \propto \prod_{d=1}^{D} \prod_{n=1}^{N_d} \beta_{k,w_{d,n}}^{\mathbb{I}[z_{d,n}=k]} \propto \text{Mult}\left(\sum_{d=1}^{D} \boldsymbol{b}_{d,k} \,\middle|\, \sum_{d=1}^{D} N_{d,k},\, \boldsymbol{\beta}_k \right)$$

where $\{\boldsymbol{w}_{n,d}\}$ are all of the words in document d and $\{\boldsymbol{z}_{n,d}\}$ are all of their assignments. Let N_d be the number of words in document d. The count vectors $\boldsymbol{b}_{d,k,v}$ and $N_{d,k}$ count the number of occurrences of word v in document d assigned to topic k and the number of occurrences of the topic k in document d, respectively:

$$\boldsymbol{b}_{d,k,v} = \sum_{n=1}^{N_d} \mathbb{I}[w_{d,n} = v]\mathbb{I}[z_{d,n} = k], \tag{11}$$

$$N_{d,k} = \sum_{n=1}^{N_d} \mathbb{I}[z_{d,n} = k].$$

We transform the word probability vectors such that $\boldsymbol{\beta}_k \equiv \pi_{\mathsf{SB}}(\boldsymbol{\psi}_k)$, introduce auxiliary variables $\boldsymbol{\omega}_k$, and set a Gaussian prior $\boldsymbol{\psi}_k \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on the stick-breaking parameters $\boldsymbol{\psi}$. Then the posterior over $\boldsymbol{\psi}$ given the counts $\{\boldsymbol{b}_d\}$ is given by the Gaussian

$$p(\boldsymbol{\psi}_k \,|\, \{\boldsymbol{b}_{d,k}\}, \{\boldsymbol{z}_d\}, \boldsymbol{\omega}_k, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \mathcal{N}\left(\boldsymbol{\psi}_k \,|\, \boldsymbol{\Omega}_k^{-1} \cdot \kappa\left(\sum_{d=1}^{D} \boldsymbol{b}_{d,k}\right), \boldsymbol{\Omega}_k^{-1}\right) \mathcal{N}(\boldsymbol{\psi}_k \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{12}$$

Dynamic Stick-Breaking Topic Model Let $t(d) \in \mathbb{N}$ denote the discrete time index of document d and $\boldsymbol{\beta}_{t,k} \in [0,1]^V$ denote the word probability vector of topic k at time t. Then we can define the following dynamical system model

$$\begin{aligned} \boldsymbol{\psi}_{t,k} &\sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\psi}_{t-1,k},\, \boldsymbol{B}\boldsymbol{B}^{\mathsf{T}}) \\ \boldsymbol{\beta}_{t,k} &\equiv \pi_{\mathsf{SB}}(\boldsymbol{\psi}_{t,k}) \end{aligned} \tag{13}$$

where $\boldsymbol{u}_{t,k}$ is a latent state of topic k at time t. Then the likelihood associated with latent state vectors $\{\boldsymbol{u}_{t,k}\}$ given the word-topic assignments $\{z_{nd}\}$ is given by the diagonal Gaussian potential

$$p(\boldsymbol{b}_{d,k} \,|\, \boldsymbol{u}_{t(d),k}, \boldsymbol{\omega}_{t(d),k}) \propto \mathcal{N}\left(\boldsymbol{\Omega}_{t(d),k}^{-1} \cdot \kappa\left(\sum_{d:t(d)=t} \boldsymbol{b}_{d,k}\right) \,\middle|\, \boldsymbol{\psi}_{t(d),k},\, \boldsymbol{\Omega}_{t(d),k}^{-1}\right). \tag{14}$$

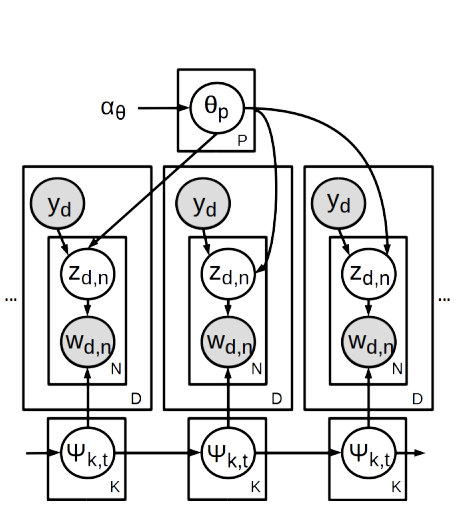
where $\boldsymbol{\Omega} \equiv \mathrm{diag}(\boldsymbol{\omega}_{t,k})$. As in equation 11, $\boldsymbol{b}_{d,k}$ counts how often each word v is assigned to topic k in document d, and the likelihood for $\boldsymbol{\psi}_{t,k}$ only depends on the documents for which $t(d) = t$. Figure 1c shows the graphical model of the stick-breaking DTM with the associated Pólya-gamma variables.

3.3 Extensions

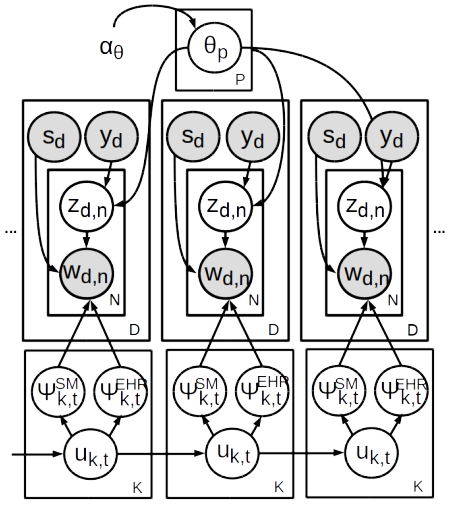
Shared Topic Proportions Among Documents In the dynamic topic model, temporal coherence arises due to the smoothness prior on $\boldsymbol{\beta}$. While this approach allows us to build temporal models from cross-sectional data, it does not use longitudinal information about whether documents are associated with the same patient when it is available.

One extension we consider is that the proportion of each disease in a patient does not change over time, that is, instead of considering a distinct document-topic vector $\boldsymbol{\theta}_d$ for each document, we have a single patient-topic vector $\boldsymbol{\theta}_p$ for each patient. However, the probability of a word given the topic—$\boldsymbol{\beta}$—will still change with time. This extension is shown in figure 2a, where we introduce the variable y_d to indicate which patient p is associated with each document d; that is, the indicator y_d selects which $\boldsymbol{\theta}_p$ to apply to document d.

Relationships between Multiple Corpora Given multiple corpora, one simple extension of the model from Section 3.2 is to posit that each disease has some canonical temporal process, but the probabilities of the terms associated with that process may vary across different corpora. For example, posts from social media may talk more about the behaviors associated with a disease, while diagnoses may focus on comorbidities. To model differences



(a) DTM with per-patient θ_p; the observed variable y_d indicates which patient the document came from.



(b) Cross-corpora extension; the observed variable s_d indicates to which corpus the document belongs.

Figure 2: Graphical Models for the DTMs in which topic proportions are shared across all notes from the same patient (2a) and DTMs that combine multiple corpora (2b). To reduce clutter, we do not include the associated Pólya-gamma variables; these are the same as in figure 1c

in term usage between corpora, we consider a dynamical system structured as

$$\begin{aligned}
\boldsymbol{u}_{t,k} &\sim \mathcal{N}(\boldsymbol{u}_{t,k} \,|\, \boldsymbol{A}\boldsymbol{u}_{t-1,k},\, \boldsymbol{B}\boldsymbol{B}^{\mathsf{T}}) \\
\epsilon_{t,k,l} &\sim \mathcal{N}(0, \sigma_l^2) \\
\boldsymbol{\psi}_{t,k,l} &\equiv \boldsymbol{u}_{t,k} + \epsilon_{t,k,l} \\
\boldsymbol{\beta}_{t,k,l} &\equiv \pi_{\mathsf{SB}}(\boldsymbol{\psi}_{t,k,l})
\end{aligned} \tag{15}$$

where now topic proportions $\boldsymbol{\beta}_{t,k,l}$ and their natural parameters $\boldsymbol{\psi}_{t,k,l}$ are associated with a specific corpus l.

Our stick-breaking construction using Pólya-gamma augmentation again renders the relevant likelihoods Gaussian: for each corpus l, the probability of the words associated with the corpus given $\boldsymbol{\psi}_{t,k,l}$ is given by

$$p(\boldsymbol{b}_{d,k} \,|\, \boldsymbol{\psi}_{t(d),k,l(d)}, \boldsymbol{\omega}_{t(d),k,l(d)}) \propto \mathcal{N}\left(\boldsymbol{\Omega}^{-1}_{t(d),k,l(d)} \cdot \kappa \left(\sum_{d:t(d)=t,l(d)=l} \boldsymbol{b}_{d,k}\right) \,\middle|\, \boldsymbol{\psi}_{t(d),k,l(d)},\, \boldsymbol{\Omega}^{-1}_{t(d),k,l(d)}\right)$$

where $l(d)$ is the corpus associated with document d, $\boldsymbol{b}_{d,k}$ is again a vector of the number of times each word v is assigned to topic k in document d from equation 11, and $\boldsymbol{\Omega} \equiv \mathrm{diag}(\boldsymbol{\omega}_{t,k})$.

Finally, the likelihood associated with the underlying temporal process $\boldsymbol{u}_{t,k}$ is simply

$$p(\boldsymbol{\psi}_{t,k,\cdot} \,|\, \boldsymbol{u}_{t,k}, \sigma_l^2) = \prod_l \mathcal{N}(\boldsymbol{\psi}_{t,k,l} \,|\, \boldsymbol{u}_{t,k}, \sigma_l^2).$$

The cross-corpora extension of the dynamic topic model is shown in figure 2b, where we explicitly show the parameters $\boldsymbol{\psi}_{t,k}^{SM}$ and $\boldsymbol{\psi}_{t,k}^{EHR}$ for just two corpora. The variable s_d indicates which source—$\boldsymbol{\psi}_{t,k}^{SM}$ or $\boldsymbol{\psi}_{t,k}^{EHR}$—should be used to model document d.

4. Inference

Given the stick-breaking dynamic topic model construction in Section 3.2, inference is straight-forward; the simplicity of inference is a key advantage of the Pólya-gamma augmentation approach. Below we summarize the inference process for the latent variables in our model: the topic proportions $\boldsymbol{\theta}_d$, the topic assignments $\{z_{nd}\}$, the topic parameters $\boldsymbol{u}$ (which can be deterministically converted into the topic proportions $\boldsymbol{\beta} = \pi_{\mathsf{SB}}(\boldsymbol{u})$), and the augmentation variables $\boldsymbol{\omega}$. The variables $\boldsymbol{\theta}$, $\{z_{nd}\}$, and $\boldsymbol{\omega}$ are resampled using Gibbs sampling, and $\boldsymbol{u}$ is resampled using a Gaussian linear dynamical system.

4.1 Resampling Document-Specific Parameters $\{z_{n,d}\}$ and $\{\boldsymbol{\theta}_d\}$

The word-topic assignments $\{z_{n,d}\}$ are resampled exactly as in the Gibbs sampler for LDA:

$$z_{nd} \sim \text{Mult}(\{\boldsymbol{\beta}_{k,v(w_{nd})}\boldsymbol{\theta}_{d,k}\})$$

where $v(w_{n,d})$ is the word associated with the token $w_{n,d}$. Likewise, the topic proportions $\boldsymbol{\theta}_d$ are also sampled exactly as in LDA:

$$\boldsymbol{\theta}_d \sim \text{Dir}(\alpha_\theta + \boldsymbol{N}_d),$$

where $\boldsymbol{N}_d$ is the vector of counts with $N_{dk} = \sum_{z_{nd}\in d} \mathbb{I}(z_{nd} = k)$. If we are sampling topic proportions per patient rather than per document, then we simply replace N_{dk} with $N_{pk} = \sum_{z_{nd}\in p} \mathbb{I}(z_{nd} = k)$, the number of times that a topic has been observed with each patient.

4.2 Resampling Topic Parameters

In the static LDA case, we can resample the natural parameters $\boldsymbol{\psi}$ from the Gaussian distribution given equation 12. In the dynamic case, we must incorporate the linear dynamical system prior.

Resampling $\boldsymbol{\psi}$: Dynamic Topic Model The formulas in equation 13 describe a linear Gaussian system, and the likelihoods in equation 14 are also Gaussian, and thus inference on $\boldsymbol{u}$ can be performed using off-the-shelf algorithms for linear dynamical systems. For completeness, we write the forward-filtering backward-sampling equations here, setting $\boldsymbol{A}$ from equation 13 to be the identity $\boldsymbol{I}$ and $\boldsymbol{B} = \mathsf{diag}(\sigma_n \dots \sigma_n)$. Define the covariance of the random walk $\boldsymbol{\Sigma} \equiv \boldsymbol{B}\boldsymbol{B}^{\mathsf{T}} = \mathsf{diag}(\sigma_n^2 \dots \sigma_n^2)$. For each topic k, we first compute the mean $q_{t,k}$ and variance $Q_{t,k}$ of the $\boldsymbol{\psi}_k$ in the forward pass:

$$\begin{aligned} q_{t,k} &= q_{t-1,k} + (Q_{t-1,k} + \boldsymbol{\Sigma})(Q_{t-1,k} + \boldsymbol{\Sigma} + \boldsymbol{\Omega}_{t(d),k}^{-1})^{-1}(y_{t,k} - q_{t-1,k}) \\ Q_{t,k} &= (\boldsymbol{I} - (Q_{t-1,k} + \boldsymbol{\Sigma})(Q_{t-1,k} + \boldsymbol{\Sigma} + \boldsymbol{\Omega}_{t(d),k}^{-1})^{-1})(Q_{t-1,k} + \boldsymbol{\Sigma}) \end{aligned} \tag{16}$$

where we start with some q_1 and Q_1 as the prior mean and variance of $\boldsymbol{\psi}_{t=1,k}$, $\boldsymbol{\Omega}^{-1}_{t(d),k}$ is computed from the auxiliary variables according to equation 14, and we use $y_{t,k} \equiv \boldsymbol{\Omega}^{-1}_{t(d),k} \cdot \kappa(\sum_{d:t(d)=t} \boldsymbol{b}_{d,k})$. Importantly, if the initial covariance Q_1 is diagonal, then because the transition covariance $\boldsymbol{\Sigma}$ and the likelihood covariance $\boldsymbol{\Omega}$ are also diagonal, the covariance $Q_{t,k}$ remains diagonal for all times t. Thus the updates in equation 16 can be computed in time linear in the size of the vocabulary $|V|$.

Similarly, the backward sampling pass can be efficiently computed by sampling $\boldsymbol{\psi}_{T,k} \sim \mathcal{N}(q_{T,k}, Q_{T,k})$ and then recursively sampling $\boldsymbol{\psi}_{t,k} \sim \mathcal{N}(q'_{t,k}, Q'_{T,k})$ where the mean $q'_{t,k}$ and variance $Q'_{T,k}$ are given by

$$\begin{aligned} q'_{t,k} &= q_{t,k} + Q_{t,k}(Q_{t,k} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{\psi}_{t+1,k} - q_{t,k}) \\ Q'_{t,k} &= (\boldsymbol{I} - Q_{t,k}(Q_{t,k} + \boldsymbol{\Sigma})^{-1})Q_{t,k} \end{aligned} \tag{17}$$

Resampling $\boldsymbol{u}$, $\boldsymbol{\psi}$: Cross-Corpora Dynamic Topic Model In the cross-corpora dynamic topic model from section 3.3, we have separate variables $\boldsymbol{u}_{t,k}$ describing the underlying dynamical system and natural parameters $\boldsymbol{\psi}_{t,k,l}$ for each corpus. Conditioned on $\boldsymbol{u}_{t,k}$, the distribution over the parameters for $\boldsymbol{\psi}_{t,k,l}$ for each time t are independent. They can be computed using equation 12 for the static LDA case and substituting the appropriate mean and variance:

$$p(\boldsymbol{\psi}_{t,k,l} \,|\, \{\boldsymbol{z}_d\}, \boldsymbol{\omega}_k \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \mathcal{N}\left(\boldsymbol{\psi}_k \,|\, \boldsymbol{\Omega}^{-1}_{t,k,l} \cdot \kappa\left(\sum_{d \in t,l}^{D} \boldsymbol{b}_d\right), \boldsymbol{\Omega}^{-1}_k\right) \mathcal{N}(\boldsymbol{\psi}_{t,k,l} \,|\, \boldsymbol{u}_{t,k}, \boldsymbol{\Sigma}_l)$$

where $\boldsymbol{\Sigma}_l$ is the diagonal covariance $\mathsf{diag}(\sigma_l^2 \ldots \sigma_l^2)$ from equation 15 and $\boldsymbol{b}_d$ sums over the word counts for topic k at time t in corpus l in document d.

Conditioned on the topic proportions $\boldsymbol{\psi}_{t,k,l}$, the evolving terms $\boldsymbol{u}_{t,k}$ can be resampled using a linear dynamical system with $\boldsymbol{\psi}_{t,k,l}$ as the emissions.

Resampling $\boldsymbol{\omega}$ In both the cross-corpora and the standard dynamic topic models, we achieve Gaussian likelihoods by augmenting the model with Pólya-gamma distributed variables $\boldsymbol{\omega}_{t,k}$ or $\boldsymbol{\omega}_{t,k,l}$ respectively. The posterior distributions of these variables are given by

$$\boldsymbol{\omega}_{t,k} | \boldsymbol{u}_{t,k}, \sim \mathrm{PG}(\boldsymbol{N}_{t,k}, \boldsymbol{u}_{t,k})$$

where $\boldsymbol{N}_{t,k}$ is a vector of how often each word appeared in all documents at time t that were assigned to topic k. In the cross-corpora case, this becomes

$$\boldsymbol{\omega}_{t,k,l} | \psi_{t,k,l}, \sim \mathrm{PG}(\boldsymbol{N}_{t,k,l}, \boldsymbol{\psi}_{t,k,l}).$$

5. Application to Learning Trajectories in Autism Spectrum Disorders

5.1 Data Description

Electronic Health Records We analyze the ICD-9CM diagnostic codes from 13,435 patients with at least one ICD-9CM code for autism spectrum disorder (299.0, 299.8, 299.9) from the Boston Children's Hospital. The Institutional Review Boards of Boston Children's

Hospital, Harvard Medical School, and the Harvard Paulson School of Engineering and Applied Sciences reviewed this study and approved it as not-human subjects research.

Each ICD-9CM code was converted into a concept unique identifier (CUIs) using the UMLS (Bodenreider, 2004) and filtered for the semantic type "Disease or Syndrome." For each code, we computed the age of the patient given the patient's birth date and the date associated with the visit that produced the code. As current evidence (e.g. Stoner et al. (2014)) suggests that ASD develops from birth, we used the age of the child as the time index for the ICD-9CM code.

To form documents, we grouped all codes associated with a patient for each year of age between the ages 0 and 15 into a "document." For example, if a patient had three visits that generated a total of ten ICD9-CM codes between ages one and two, and two more visits that generated a total of five ICD9-CM codes between ages two and three, then that patient would be associated with two documents: one at time index "age 1," with ten codes, and one at time index "age 2," with five codes. Grouping all diagnostic codes from a year into one document smoothed over variations due to visits to specialties that focused on different aspects of the child's care. This processing procedure resulted in 63,941 documents with an average of 5.3 CUIs each and 7,037 unique CUIs.

Social Media We scraped all subforums of the websites `www.asd-forum.org.uk`, `www.autismweb.com`, and `www.asdfriendly.org`, resulting in 664,954 posts from 80,927 threads. An example post is given in Appendix A.1. The forum posts contained the date of posting but not the child's date of birth; thus additional processing was required to determine the age of the child—and thus the time-index—for the documents. Regular expressions (see Appendix A.2) were used to extract ages from the posts, and posts with multiple ages were excluded. This procedure resulted in 13,743 posts with a single mention of age. Approximately 1,000 of these posts were hand-checked for accuracy; the regular expressions were adjusted to avoid any errors that were discovered in the hand-checked posts.

We filtered for patients between 0 and 15 years of age, and as with the electronic health records, we combined all the posts written about the same patient with the same age into one document to smooth over variations due to the caregiver's particular concerns at the time. This processing resulted in a data set of 5,461 documents (each containing possibly multiple posts written in the same year) by 2,391 unique users.

Clinically-relevant terms were extracted from these posts by finding terms that matched the consumer health vocabulary (Zeng, 2015), which has mappings into the UMLS CUIs. A trie was used to quickly match terms to the dictionary of words, and only terms with the semantic type "Disease or Syndrome" were included. The average number of CUIs per document was 1.8. Of the 7,372 CUIs across the EHR and SM data sets, 284 were unique to the forum posts and 2,407 were unique to the EHR codes.

5.2 Methods

Models We considered three variants of dynamic topic models:

- *SB-DTM-*$\boldsymbol{\theta}_d$ The stick-breaking DTM from Section 3.2.

- *SB-DTM-$\boldsymbol{\theta}_p$* The stick-breaking DTM in which we assume that distribution over diseases in each patient remains constant over time, as described in Section 3.3.

- *SB-ccDTM* The stick-breaking DTM in which the EHR and SM corpora are modeled as having distinct topics with shared underlying dynamics, as described in Section 3.3.

These variants were compared to two versions of LDA: in the first version, LDA-K was trained with K topics that did not evolve over time. LDA-K15 was trained with $15K$ topics, accounting for the fact that the dynamic topic model could have a different topic for each year in ages 0 to 15.[1]

Evaluations Our first evaluation metric was simply predictive log-likelihoods. We randomly held-out 10% percent of the words from 10% percent of the documents. Once the model was trained, we had a value of the topic proportions $\boldsymbol{\theta}_d$ for every document d. Thus, probability of a held-out word w_{nd} was given by

$$p(w_{nd} \,|\, \theta_d, \boldsymbol{\beta}) = \sum_z p(w_{nd} \,|\, z, \boldsymbol{\beta}_{z,t(d)}) p(z \,|\, \theta_d)$$

Our second evaluation metric simulated the more clinically relevant task of stratifying patient risk for various future outcomes. For this evaluation, we considered only patients with at least one document during early childhood—under the age of five—and one document from later childhood—over the age of seven. For 10% of these patients, we held out *all* the documents for after the child was six years old. The documents from when the child was five years old or younger were included in the DTM training. Following training, we computed the average document-topic proportions $\boldsymbol{\theta}_p$ for each patient as

$$\boldsymbol{\theta}_p = \frac{1}{N_p^{\leq 5}} \sum_{ds.t.t(d) \leq 5} \boldsymbol{\theta}_d$$

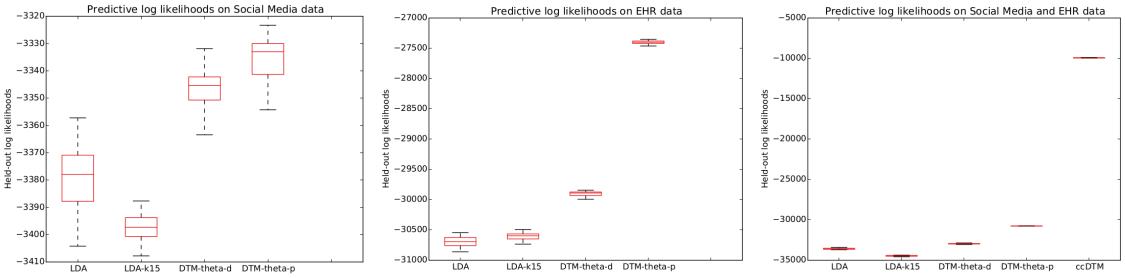
where $N_p^{\leq 5}$ is the number of documents associated with patient p where $t(d) \leq 5$. This averaging corresponds to the assumption that the patient's disease proportions do not change over time; note that in the shared-proportions DTM from Section 3.3, we can simply used the learned $\boldsymbol{\theta}_p$.

Given a patient-topic vector $\boldsymbol{\theta}_p$, we can compute the likelihood of the future, *unseen* notes

$$p(w_{n,d} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\beta}) = \sum_{z, ds.t.d \geq 7} p(w_{n,d} \,|\, z, \boldsymbol{\beta}_{z,t(d)}) p(z \,|\, \boldsymbol{\theta}_p)$$

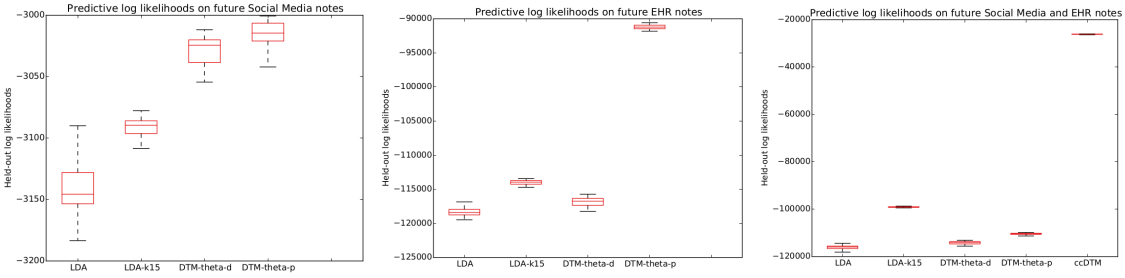
If our temporal models were capturing time-varying patterns in disease processes, we would expect our model to better predict the content of future documents than a static model.

1. We also ran tests using the C implementation of dynamic topic models available at `https://github.com/blei-lab/dtm` but were unable to achieve satisfying likelihoods with several parameter settings.



(a) SM (b) EHR (c) SM and EHR

Figure 3: Boxplots of held-out test likelihoods for the different models on SM data alone, EHR data alone, and both data sets combined. Across all versions, the dynamic models have higher predictive performance.



(a) SM (b) EHR (c) SM and EHR

Figure 4: Boxplots of predictions of future patient notes on SM data alone, EHR data alone, and both data sets combined. Models which are trained with the assumption that topic proportions $\boldsymbol{\theta}_p$ for each patient remain constant over time do best in the individual data sets, and the transfer learning in the combined case has the best predictive performance.

5.3 Results and Analysis

We completed 10 runs each of LDA, LDA-K15, the standard DTM, the SB-DTM-θ_d, the SB-DTM-θ_p, and the SB-ccDTM. We completed runs on the EHR data alone, the SM data alone, and the SM and EHR data combined. The results of LDA were used to initialize the dynamic topic models, and the results of basic DTM were used to initialize the ccDTM. Preliminary tests of 300 iterations showed that the samplers mixed by around 50 iterations (see figure 9 in appendix B for an example plot); in the results below each sampler was run for 100 iterations. The transition noise parameter in the linear dynamical system was set to $\sigma_n^2 = 0.1$, the cross-corpora noise parameter in SB-ccDTM was set to $\sigma_f^2 = 1$, and the number of topics K was set to 10 based on initial parameter exploration of $K = 5, 10, 15$.

Predictive Performance: Held-out Data Figure 3 shows the held-out test likelihoods for the SM, EHR, and combined cohorts, respectively, for $K = 15$. We see that the dynamic models outperform the static models, including an LDA model with as many topics as the DTM. Indeed, LDA-K15 has the lowest overall predictive likelihoods, suggesting that it may be overfitting. Incorporating links between notes from the same patient (DTM-$\boldsymbol{\theta}_p$)

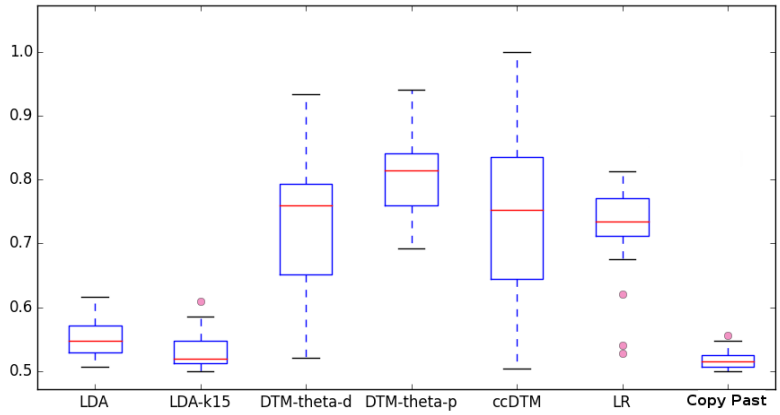


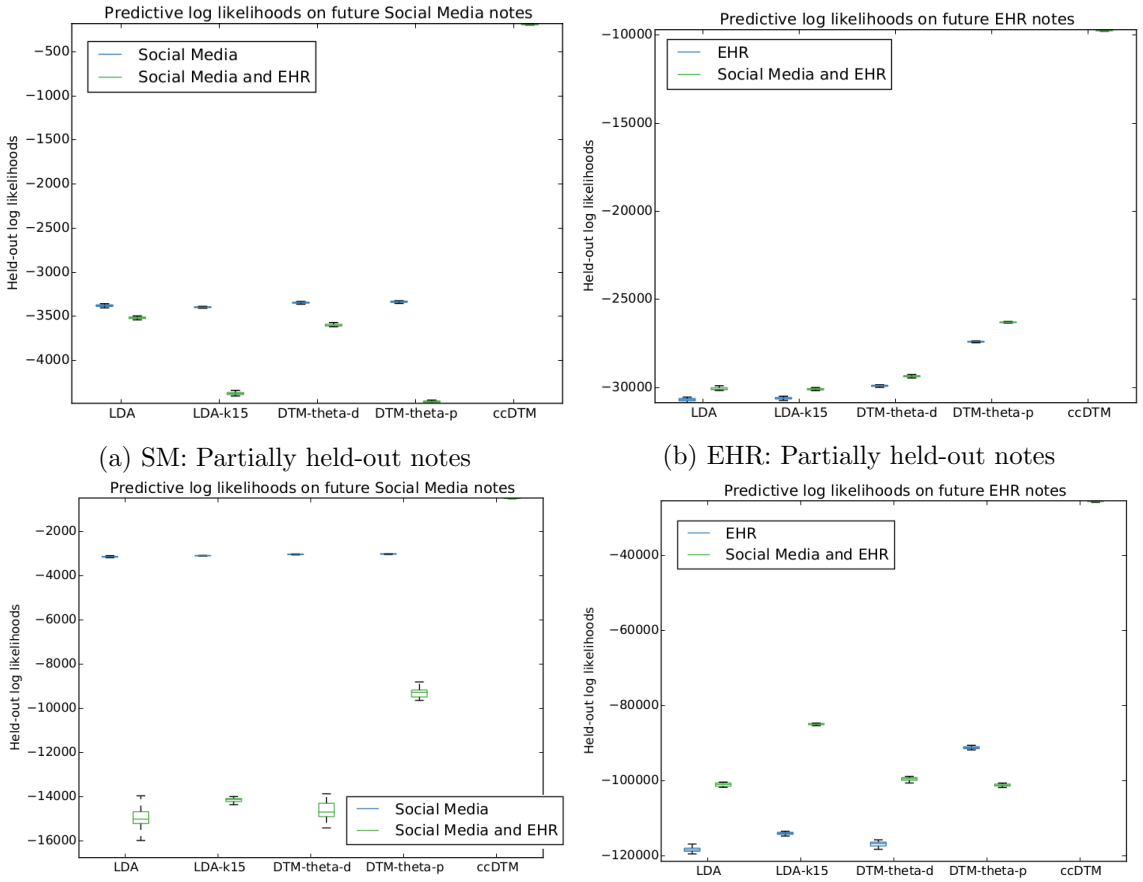
Figure 5: Boxplots of AUCs for predicting future patient conditions for conditions that occurred in at least 10% of the patients. Even without explicitly trying to optimize future predictions, the DTM-based approaches are comparable to—or better than—a discriminative baseline such as logistic regression.

improves prediction quality in both the individual and combined data sets, and the added flexibility of the cross collection ccDTM model further improves prediction accuracy in the combined data set.

Predicting Future notes Figure 4 shows the held-out test likelihoods for the content of *all* patients notes associated with age seven and above given all the notes from that patient under the age of five. Predicting the content of an *entirely* held-out note is much harder than predicting the missing contents of a partially held-out note. We see that training the models with the assumption that topic proportions stay constant—as in the DTM-$\boldsymbol{\theta}_p$ model—results in the best predictive performance on these entirely held-out notes in both data sets. In the combined data set, the ccDTM model, which also allows for transfer learning between the SM and EHR data sets, achieves the highest predictive likelihoods.

Figure 5 shows AUCs for the same task of predicting the contents of future notes given current ones. We see that the DTM-based models again perform better than their static counterparts because they are able to imagine what future diseases may occur (the boxplots are over all CUIs with at least 10% prevalance the future notes). The DTM model which takes advantage of the links between patients performs the best, better than the logistic regression discriminitive baseline. Finally, we see that simply assuming that a patient's past condition will continue into the future (copy past) does not produce high AUCs; both the DTM and the logistic regression are learning meaningful predictive relationships.

Transfer Learning between EHR and Social Media The previous analyses showed that our dynamic models better predict held-out and future patient data than a static model. It is also interesting to test whether combining the two data sets increases predictive performance on data from each of the individual data sets, that is, for the some held-out



(a) SM: Partially held-out notes

(b) EHR: Partially held-out notes

(c) SM: Entirely held-out future notes

(d) EHR: Entirely held-out future notes

Figure 6: Boxplots comparing of predictions of randomly held-out data and future notes for each cohort vs. the combined cohort. In general, transfer is positive for EHRs and negative for SM; however, the flexibility of the cross-corpora DTM results in positive transfer in all scenarios (tiny bar at the top-right in all the plots).

EHR data, is there a benefit to training on EHR and SM data rather than training on EHR data alone? Likewise, for some held-out SM data, does adding EHR data into the training set benefit predictive performance? (Note that there is no reason, a priori, to assume that combining collections will be beneficial.)

The boxplots in figure 6 show the results of this test for both randomly held-out data and for predicting entire future notes. The blue boxplots correspond to training only on the target data set, and the green boxplots correspond to training on the combined data set. In the EHR cohort, the transfer is positive in almost all cases (the green boxplots are higher than their corresponding blue boxplots), even among models such as standard LDA. The opposite is true in the SM cohort: training on the combined set decreases predictive accuracy among the flat models, likely because the SM data set had many fewer documents than the EHR data set. However, in all cases, we observe that cross-collection DTM (ccDTM)—

whose hierarchy allows greater flexibility in how information is shared across the two data sources—has the highest predictive performance.

Computational Time The static LDA models had the fastest wall-clock times, with a five-topic LDA model on the full corpus taking 0.279 seconds per iteration and the larger LDA-15 taking 0.414 seconds per iteration. The standard DTM-$\boldsymbol{\theta}_d$ took 2.00 seconds per iteration, and adding the patient links in the DTM-$\boldsymbol{\theta}_p$ increased the per iteration runtime to 2.14 seconds per iteration. Interestingly, the ccDTM required only 0.953 seconds per iteration, because the forward-backward pass over the $\boldsymbol{u}$ variables only had two emissions—the $\boldsymbol{\psi}$ from each corpus—rather than inputs from all of the documents.

Qualitative Examination of Topics: Electronic Health Records We show the top-4 words for the EHR-only θ-p DTM in table 1. (We choose a small K for brevity, larger K have similar and additional patterns.) Topic 0 corresponds to the trajectory of patients with ASD who also have Down's syndrome. ASD and Down's syndrome are known to be comorbid with each other (Kent et al., 1999; Rasmussen et al., 2001). Expressive disorder, a feature of both ASD and Down's syndrome, shows up in the top-4 list as children are learning language at age 2; later the top-4 list is dominated by clinical features such as infections. The overall prevalence of infection-related terms is consistent with associations of immunodeficiency with both Down's syndrome (Ram and Chinen, 2011) and ASD (Gupta et al., 2010), including increased ear infections specifically (Konstantareas and Homatidis, 1987b). Children with Down's syndrome are more likely to have a variety of abnormal ocular features such as myopia (Shapiro and France, 1985) and abnormalities of the ear such as eustachian tube dysfunction (Pueschel, 1990; Shott et al., 2001). Sleep apnea is also common in children with Down's syndrome (Marcus et al., 1991).

Topic 1 corresponds to children with ASD who go on to develop psychiatric disorders, and is very similar to the psychiatric subgroup reported by Doshi-Velez et al. (2013). As expected, there is a progression from ADHD at age 4, anxiety and conduct disorders at age 10, to episodic mood disorders at age 15 (other prevalent, but not top-4 terms at age 15 included depressive disorder and childhood psychoses). Psychiatric disorders are commonly reported among higher functioning children with ASD (Gillott et al., 2001; DeLong and Dwyer, 1988), and the progression of diagnoses makes sense because clinicians will usually avoid giving a young child a diagnosis for a severe psychiatric illness.

Topic 2 contains a combination of intellectual disability and epilepsy. It is similar to neurological subgroup reported by Doshi-Velez et al. (2013). Epilepsy is a common comorbidity of autism (Sherr, 2003a; Mouridsen SE, 1999), affecting close to 20% of children with ASD. Sherr (2003b) suggest that these three disorders—epilepsy, intellectual disability, and ASD—are linked through the ARX gene. Laumonnier et al. (2004) find common genes between ASD and intellectual disability, and Sharp et al. (2008) report genomic underpinnings for epilepsy and intellectual disability. Again, a young child is less likely to be given a diagnosis of intellectual disability—it appears in our top-4 list at age 4—but other signs, such as symbolic dysfunction and developmental delays are noted from infancy.

Topic 3 tracks the progression of children with ASD and cerebral palsy. There are known correlations between cerebral palsy and infantile autism (Surén et al., 2012; Talkowski et al., 2012); early infections (seen at age 0) have also been associated with both cerebral palsy and autism spectrum disorders (Konstantareas and Homatidis, 1987a; Rosenhall et al., 1999).

Table 1: Top words from Dynamic Topic Model trained only on Electronic Health Records.

	Year 0	Year 2	Year 4	Year 10	Year 15
Topic 0	Otitis Media, Down Syndrome, Acute upper respiratory infection, Unspecified viral infection	Expressive Language Disorder, Otitis Media, Down Syndrome, Chronic serous otitis media	Otitis Media, Expressive Language Disorder, Down Syndrome, Eustachian tube disorder	Down Syndrome, Eustachian tube disorder, Sensorineural Hearing Loss, Otitis Media	Down Syndrome, Eustachian tube disorder, Sleep Apnea, Myopia
Topic 1	Acute bronchiolitis, Asthma Redundant prepuce and phimosis, Chronic maxillary sinusitis	Other specified pervasive developmental disorders, Asthma, Urea Cycle Disorders, Autistic Disorder	Other specified pervasive developmental disorders, Attention deficit hyperactivity disorder, Autistic Disorder, Developmental delay (disorder)	Attention deficit hyperactivity disorder, Other specified pervasive developmental disorders,, Anxiety state, Conduct Disorder	Attention deficit hyperactivity disorder, Other specified pervasive developmental disorders, Anxiety state, episodic mood disorders
Topic 2	Other specified delays in development, Mixed development disorder, Viral and chlamydial infection, Developmental delay (disorder), Symbolic dysfunction	Infantile autism, Symbolic dysfunction, Developmental delay (disorder), Other specified delays in development	Symbolic dysfunction, Infantile autism, Unspecified intellectual disabilities, Epilepsy	Infantile autism, Unspecified intellectual disabilities, Epilepsy, Symbolic dysfunction	Infantile autism, Unspecified intellectual disabilities, Epilepsy, unspecified, Generalized convulsive epilepsy,
Topic 3	Infantile cerebral palsy, Gastroesophageal reflux disease, Chronic respiratory disease in perinatal period, Deglutition Disorders	Infantile cerebral palsy, Quadriplegic Infantile Cerebral Palsy, Diplegic Infantile Cerebral Palsy, Deglutition Disorders	Infantile cerebral palsy, Quadriplegic Infantile Cerebral Palsy, Diplegic Infantile Cerebral Palsy, Deglutition Disorders	Quadraplegic Infantile Cerebral Palsy, Infantile cerebral palsy, allergic rhinitis, hay fever	Quadraplegic Infantile Cerebral Palsy, Infantile cerebral palsy, Hemiplegic cerebral palsy, Gastroesophageal reflux disease
Topic 4	Gastroesophageal reflux disease, Atrial septal defect within oval fossa, Hypoplastic Left Heart Syndrome, DiGeorge Syndrome	Gastroesophageal reflux disease, Deglutition Disorders, Asthma, Failure to Thrive	Gastroesophageal reflux disease, Muscle, ligament and fascia disorders, Developmental Coordination Disorder, Deglutition Disorders	Gastroesophageal reflux disease, Hypogamma globulinemia, Asthma, Hematological Disease	Hypogamma globulinemia, Gastroesophageal reflux disease, Adjustment Disorder With Mixed Anxiety and Depression, Hyperopia

Children with cerebral palsy are known to have difficulty swallowing (Sochaniwskyj et al., 1986) and reflux (Reyes et al., 1993). Horvath et al. (1999) also note an association between ASD and a number of gastrointestinal symptoms, including increased reflux.

Finally, topic 4 initially contains a variety of more severe multi-system disorders. Many are congenital anomalies (e.g. DiGeorge Syndrome and septal defects), which are more prevalent in ASD (Wier et al., 2006). It makes sense to see "failure to thrive"—usually diagnosed in early childhood—as one of the top diagnoses in this topic of severe illnesses. The later terms contain features common in ASD (GI symptoms, immunodeficiency) seen in earlier topics but without the associated Down's syndrome or cerebral palsy. This topic is somewhat reminiscent of the multi-system subgroup in Doshi-Velez et al. (2013). More broadly, analyzing the same data set, we recover topics that resemble the subgroups discovered in Doshi-Velez et al. (2013) with the addition of specific trajectories for patients with ASD and Down's syndrome and ASD and cerebral palsy.

Qualitative Examination of Topics: Social Media Table 2 shows a similar table for the θ_p DTM trained on the social media data alone. Even after filtering for only signs and symptoms, the extracted terms from the forum posts tend to focus more the symptoms of the child's ASD rather than other comorbid conditions. [2] Topic 3 seems to correspond to the most "traditional" ASD trajectory, with speech delays and tantrums early on. Emotional distress is a constant, and we see that bullying makes the top-4 list at age 10. Children with ASD are both more likely to bully (Montes and Halterman, 2007; Van Roekel et al., 2010) and be bullied (Lee et al., 2008), especially as they reach later grade school and early middle school years.

In general, terms such as tantrums and mental suffering are common in many of the topics. For example, topic 0 follows the trajectory of children with stereotypies (tic disorder, apraxias) common in ASD (Goldman et al., 2009). Pagnamenta et al. (2010) suggest genetic commonalities between ASD and dyslexia, and Gillon and Moriarty (2007) note that children with speech apraxias are also at higher risk for dyslexia. However, there also exists a parallel set of terms starting with mental suffering starting at age 2 and ending with psychiatric problem at age 15. Even if some of the mental suffering terms are a mistaken reference to the challenges experienced by the caregiver, rather than the child, we can still say that forums generally contain more language pretaining to mental health.

Topic 1 describes emotional distress, including nightmares (while nightmares are not reported as common in the clinical literature, sleep disorders are very common and it may be that parents attribute sleep disorders to nightmares (Gail Williams et al., 2004)), as well as reactions that children may have to stress—temper tantrums and aggressive behaviors. These terms turn to phobic anxiety at later ages. We conjecture that this topic is the care-giver analog of EHR Topic 1 above, which followed the trajectories of patients with psychiatric disorders.

Like Topic 2 in the EHR, Topic 2 here describes developmental delays and epilepsy. However, we see abstract thought disorder rather than intellectual disability as well as symptoms such as staring. At age 15, emotional distress again makes the top-4 list, suggesting that most children with ASD face challenges as they grow older and interact more

2. We were not able incorporate clinical notes in this study, but it is possible that the clinical note would also tip the balance toward terms describing the patient's ASD rather than other comorbidities.

with society. Topic 4 also has some psychiatric disorders, including aggressive behavior turning into emotional distress, bullying, and depression as the child ages.

Table 2: Top words from Dynamic Topic Model trained on only Social Media

	Year 0	Year 2	Year 4	Year 10	Year 15
Topic 0	Infection, Apraxias, Developmental delay (disorder), Autistic Disorder	Autistic Disorder, Infection, Apraxias, Mental Suffering,	Autistic Disorder, Apraxias, Tic disorder, Mental Suffering	Autistic Disorder, Dyslexia, Apraxias, Mental Suffering	Autistic Disorder, Apraxias, Tic disorder, Psychiatric problem
Topic 1	Autistic Disorder, Emotional distress, Abstract thought disorder, Temper tantrum	Autistic Disorder, Emotional distress, Abstract thought disorder, Aggressive behavior	Autistic Disorder, Emotional distress, Abstract thought disorder, Aggressive behavior	Autistic Disorder, Emotional distress, Temper tantrum, Aggressive behavior	Autistic Disorder, Emotional distress, Phobic anxiety disorder, Nightmares
Topic 2	Autistic Disorder, Abstract thought disorder, Temper tantrum, Staring	Autistic Disorder, Temper tantrum, Abstract thought disorder, Epilepsy	Autistic Disorder, Abstract thought disorder, Temper tantrum, Staring	Autistic Disorder, Abstract thought disorder, Temper tantrum, Epilepsy	Autistic Disorder, Abstract thought disorder, Asperger Syndrome, Emotional distress
Topic 3	Autistic Disorder, Speech Delay, Emotional distress, Temper tantrum	Autistic Disorder, Speech Delay, Emotional distress, Temper tantrum	Autistic Disorder, Emotional distress, Psychiatric problem, Speech Delay	Autistic Disorder, Emotional distress, Bullying, Asperger Syndrome	Autistic Disorder, Mental Suffering, Emotional distress, Apraxias
Topic 4	Autistic Disorder, Aggressive behavior, Nightmares, Apraxias	Autistic Disorder, Forgetting, Aggressive behavior, Mental Suffering	Autistic Disorder, Emotional distress, Temper tantrum, Confusion	Aggressive behavior, Emotional distress, Violent, Bullying, Forgetting	Emotional distress, Mental Depression, Violent, Mental Suffering

Qualitative Examination of Topics: Cross-corpora model Finally, we show the matching topics of the cross-corpora model in tables 3 and 4, as well as the overall proportions of each topic in figure 7. Again, we limit ourselves to a smaller topic model and show only a few top words, but we emphasize that in a clinical application these choices can be expanded and each topic examined in significantly more detail. What is most interesting for our purposes is the cross-corpora DTM allows us to see where top words in the corpora match, and where they do not.

Overall, the topics are closer to the EHR topics—likely a reflection of the fact that we had more EHR data. For example, topic 0 appears to be epilepsy topic (with pervasive developmental disorders replacing intellectual disability as a topic term, but reflecting a similar set of conditions). Epilepsy-related terms are also present in the social media version of the topic; however, we also see ADHD—also comorbid with epilepsy (Surén et al., 2012; Dunn et al., 2003)—present in both topics, likely because ADHD is commonly discussed on forums. We also dental caries, which are also associated with epilepsy (Anjomshoaa et al.,

2009), in the social media version of the topic. Such dental terms would not occur as often in the clinical records because children see their dentists outside the hospital system.

Topic 1 contains several psychiatric disorders with increasing severity (especially prominent in the EHR version of the topic). These show up as more general emotional distress and mental suffering in the forum topic. While most of the topics are present in similar relative proportions in both corpora (figure 7), topic 1 is the most common topic in the social media source and the least common topic in the electronic health records. We posit this difference may be because caregivers in general may be more focused on the mental health of their children (as seen in the social media-only topics), while the EHRs contain a range of specialties seen by the patient and perhaps disproportionately little about their mental health.

Topic 2 contains many infections, in both the social media and the EHR, which are consistent with the immunodeficiency-related topics discovered from EHR alone. Interestingly, asthma, an autoimmune disease, also appears in this topic; Becker (2007) posits that some ASDs, asthma, and inflammation may have a common autoimmune component. Doshi-Velez et al. (2013) also found a subgroup enriched for asthma. Obesity, associated with asthma (Beuther et al., 2006), also appears in this topic; here it seems that combining the sources resulted in a much clearer infections and autoimmune topic rather than the more diluted multi-system EHR topic 4.

Finally, topic 3 mirrors the cerebral palsy topic from the EHRs and topic 4 mirrors the Down's syndrome topic. In the cerebral palsy topic (topic 3), we see more differences in the topics early on. Caregivers mention temper tantrums, speech delays, and abstract thought disorder—all features consistent with ASD and cerebral palsy—early on but the term cerebral palsy does not make the top-4 list. Later the terms are more similar across the two sources. Similarly, the caregiver version of the top-4 list for the ASD and Down's syndrome topic (topic 4) includes more terms like expressive language disorder and symbolic dysfunction early on as well as stereotypic movements.

6. Related Work

Disease Progression Models Disease progression modeling is an important area in medical informatics. When biomarkers of interest are known, or disease stages have been labeled, supervised approaches can be used to predict disease stages given signs and symptoms; such supervised approaches have been applied to modeling the progression of Alzheimer's disease (Zhou et al., 2012a). Other approaches use physiological models (De Winter et al., 2006) or meta-analyses of existing literature (Ito et al., 2010) to derive disease progression models.

One of the most popular data-driven approaches to learning disease progression models is to fit a hidden Markov Model (HMM) to the observations. The states of the HMM correspond to different stages of chronic diseases, and often left-to-right HMMs are used model the fact that many disease progression processes are not reversible. Such models have been used to model disease progression in chronic kidney disease (Luo et al., 2013; Yang et al., 2014), Alzheimer's disease (Sukkar et al., 2012), aneurysm screening (Jackson et al., 2003), and flu (Fan et al., 2015). Yang et al. (2014) allow the patient to have multiple conditions at the same time, treating each patient as having a mixture of disease pathways. Luo et al. (2013) take into account irregular sampling of data.

Table 3: Top words from Dynamic Topic Model trained on both SM and EHR data.

	Year 0	Year 2	Year 4	Year 10	Year 15
EHR Topic 0	Acute upper respiratory infection, Hearing Loss, gastroenteritis, Hirschsprung's disease	Expressive Language Disorder, Developmental delay, Hearing Loss, Mixed development disorder	Expressive Language Disorder, Developmental delay, Epilepsy, Localization-related epilepsy	Epilepsy, Hearing Loss, Conduct Disorder, ADHD	Generalized tractable convulsive epilepsy, Conduct Disorder, Generalized intractable convulsive epilepsy, Epilepsy
SM Topic 0	Epilepsy, Acute upper respiratory infection, Mixed development disorder, Hemophilia B	Ehlers-Danlos Syndrome, Developmental delay, Ritual compulsion, Dental caries	Exanthema, Developmental delay, Pervasive Development Disorder, Metabolic Diseases	Hearing Loss, Mixed Conductive-Sensorineural Disorder, Hearing Loss, Dental caries	Generalized intractable convulsive epilepsy, Dental caries, ADHD, Grand Mal Status Epilepticus
EHR Topic 1	Acute bronchiolitis, Redundant prepuce and phimosis, Common Cold, Epilepsy	Autistic Disorder, pervasive developmental disorders, Asthma, Urea Cycle Disorders	ADHD, Autistic Disorder, speech or language disorder, Urea Cycle Disorders	ADHD, Other specified pervasive developmental disorders,, Tic disorder, Autistic Disorder	pervasive developmental disorder, ADHD, Psychotic Disorders, Depressive disorder, Emotional distress
SM Topic 1	Autistic Disorder, Emotional distress, Abstract thought disorder, Epilepsy	Autistic Disorder, Emotional distress, Mental Suffering, Aggressive behavior	Autistic Disorder, Emotional distress, Aggressive behavior, Tic disorder	Autistic Disorder, Emotional distress, Bullying, Aggressive behavior	Autistic Disorder, Emotional distress, Aggressive behavior, Mental Suffering
EHR Topic 2	Otitis Media, Atrial septal defect within oval fossa, Acute upper respiratory infection, Viral infection	Otitis Media, Asthma, Acute upper respiratory infection, Spina bifida	Otitis Media, Asthma, Spina bifida, Unspecified viral infection	Asthma, Otitis Media, Developmental delay, Obesity	Hypogamma globulinemia, Sleep Apnea, Asthma, Obesity
SM Topic 2	Acute upper respiratory infection, Atrial septal defect within oval fossa, Otitis Media, Vesicoureteral reflux,	Common Cold, Forgetting, Exanthema, Asthma	Common Cold, Forgetting, Asthma, Urinary tract infection	Exanthema, Common Cold, Obesity, Asthma	Developmental delay, Hypogamma globulinemia, Enlargement of tonsil or adenoid, Anomalous pulmonary artery

Table 4: Top words from Dynamic Topic Model trained on both SM and EHR data.

	Year 0	Year 2	Year 4	Year 10	Year 15
EHR Topic 3	Gastroesophageal reflux disease, Deglutition Disorders, Congenital Hypothyroidism, Chronic respiratory disease in perinatal period	Infantile cerebral palsy, Deglutition Disorders, Gastroesophageal reflux disease, Quadriplegic Infantile Cerebral Palsy	Infantile cerebral palsy, Quadriplegic Infantile Cerebral Palsy, Gastroesophageal reflux disease, Deglutition Disorders	Infantile cerebral palsy, Quadriplegic Infantile Cerebral Palsy, Gastroesophageal reflux disease, Diplegic Infantile Cerebral Palsy	Quadriplegic Infantile Cerebral Palsy, Gastroesophageal reflux disease, Hemiplegic cerebral palsy, Intellectual disabilities
SM Topic 3	Deglutition Disorders, Infantile cerebral palsy, Congenital Hypothyroidism, Gastroesophageal reflux disease	Speech Delay, Temper tantrum, Abstract thought disorder, Developmental delay	Abstract thought disorder, Temper tantrum, Developmental delay, Gastroesophageal reflux disease, Muscle, ligament and fascia disorders	Diplegic Infantile Cerebral Palsy, Myopia, Failure to Thrive, Other specified delays in development	Quadriplegic Infantile Cerebral Palsy, Infantile cerebral palsy, Generalized convulsive epilepsy, Other specified delays in development
EHR Topic 4	Down Syndrome, Atresia and stenosis of large intestine, Contact dermatitis, Middle ear conductive hearing loss	Down Syndrome, Infantile autism, Symbolic dysfunction, Eustachian tube disorder	Symbolic dysfunction, Infantile autism, Other specified pervasive developmental disorders,, Down Syndrome	Infantile autism, Down Syndrome, pervasive developmental disorders, Anxiety state	Infantile autism, Anxiety state, Down Syndrome, Intellectual disabilities
SM Topic 4	Down Syndrome, Middle ear conductive hearing loss, Eustachian tube disorder, Unspecified intellectual disabilities	Autistic Disorder, Infantile autism, Expressive Language Disorder, Symbolic dysfunction	Autistic Disorder, Eustachian tube disorder, Stereotypic Movement Disorder, Hay fever, Down Syndrome	Pervasive developmental disorders,, Stereotypic Movement Disorder, Hay fever, Asthma	Psychotic Disorders, Down Syndrome, Unspecified childhood psychosis, Other specified pervasive developmental disorders

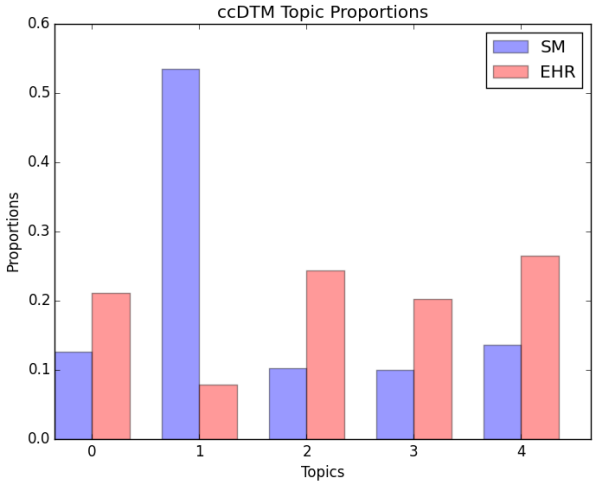


Figure 7: Overall topic popularities in both electronic health record and social media documents. Except for topic 1, most of the topics are present in similar proportions.

Others model disease progression with continuous time processes. Liu et al. (2013) model the progression of glaucoma with continuous-time HMMs, while Wang et al. (2014) use continuous-time Markov jump processes to model the progression of chronic obstructive pulmonary disease. Saeedi and Bouchard-Côté (2011) introduce gamma-exponential processes to model recurrent disease processes multiple sclerosis. These models can be adapted to incorporate individual-specific progression rates and treatment effects (Post et al., 2005).

Whether discrete or continuous time, all of these approaches involve discrete disease stages. However, often diseases evolve slowly over time. While we use a discrete time model in our work, a fundamental difference in our approach from those above is that we do not attempt to divide disease progression into stages, which might be artificial distinctions. Especially in developmental disorders, a more continuous progression model is more natural as a child's development is a continuously evolving process. In this sense, perhaps closest in spirit to our work is the work of Zhou et al. (2014), which models disease progression with a matrix factorization that is smooth in the time dimension. Che et al. (2015) embed each time point of a patient into a latent space using a deep network.

In addition to being a natural way to model smoothly evolving diseases, our smoothness assumption allows us to easily incorporate cross-sectional data as well as longitudinal data. Requiring multiple visits to derive trajectories is often one of the factors that greatly limits the amount of data that can be used from a cohort: Doshi-Velez et al. (2013) used EHRs from the same hospital as us but were limited to only 4,927 patients with many visits rather than the 13,435 patients we study here (unlike Doshi-Velez et al. (2013) and other clustering-based studies, we do also not rely on ad-hoc patient similarity functions and intensive data pre-processing). Other studies that use smoothness assumptions in similar ways are Ross et al. (2014) and Li et al. (2012). Li et al. (2012) derive trajectories and then define an HMM from cross-sectional data through temporal bootstrap method that connects patients with similar features; their approach has no underlying model but rather

relies on patient similarities to build trajectories. Ross et al. (2014) derive lung capacity trajectories in chronic obstructive pulmonary disease from a cross-sectional cohort using Gaussian processes to encourage smoothness.

Dynamic Topic Models and Dynamical Systems Several techniques exist to model the temporal evolution of topics. Wang and McCallum (2006) consider the case in which the popularity of a topic changes over time, but each topic's word proportions remain stationary. In contrast, dynamic topic models (Blei and Lafferty, 2006b; Wang et al., 2012) assume a topic's word proportions smoothly evolve over time. Dynamic topic models have been applied to applications including discovering themes in research communities, (Furukawa et al., 2015), evolving patterns in software programs (Thomas et al., 2014), and the adoption of applications by smart phone users (Chua et al., 2015).

Topic models have also been developed for modeling multiple corpora. Wang et al. (2009) model correlations between the natural parameters for multiple corpora as a Gaussian random field. Paul (2009); Paul and Girju (2009); Zhai et al. (2004) model correlations between multiple corpora through a mixture of base and corpora-specific topics. Zhang et al. (2010) model the changing popularities of topics across three corpora—blogs, news, and message boards—using evolutionary hierarchical Dirichlet processes.

There also exists a related literature on modeling text as dynamical systems. Mikolov (2012) model dependencies in text as a recurrent neural network. Belanger and Kakade (2015) model text as a Gaussian linear dynamical system. Their model is misspecified in that it attributes zero probability mass to any observation, but they note the computational convenience of modeling occurrences of words with Gaussian variables rather than multinomials. While we are not modeling sequences of words, the idea modeling trends as linear dynamical system is close in spirit to our work.

In this context, we emphasize that the models we described in Section 3 are not novel—dynamic topic models and cross-corpora topic models both have well-established literatures. However, each topic model variant above relies on its own bespoke, implementation-intensive inference techniques that are often specific to that model. By using Pólya-gamma augmentation in our inference, we are able easily explore a variety of models. Moreover, to our knowledge, the application of dynamical system models of text to characterize disease progression is novel.

Disease Models from Social Media There exists a large body of work analyzing social media for information related to diseases. Chee et al. (2011) use personal health messages to predict adverse drug events, while Wilson and Brownstein (2009); Paul et al. (2015) use social media for disease surveillance. Elhadad et al. (2014); Jha and Elhadad (2010) characterize the linguistic properties of online forum text and use it to predict the cancer stage of the patient. Coppersmith et al. (2015) describe the task of identifying patients with depression and post-traumatic stress disorder from their Twitter posts. Unlike these works, our objective is understanding disease phenotypes and disease progression from social media, not prevalence or diagnosis.

7. Discussion and Conclusions

Modeling Choices Using dynamic topic models for modeling disease progression offers several advantages over more traditional clustering and HMM-based approaches. We do not require patients to belong to a single cluster or health state; they may have multiple disease processes varying in intensity over time, and each disease process is a smoothly varying, rather than discrete, structure. Because we can combine longitudinal and cross-sectional data, we can take advantage of much larger cohorts. Unlike clustering approaches, no ad-hoc patient similarity metrics are required, and unlike HMM-based approaches, we do not need to perform inference about what may have happened to patients in the gaps between visits.

Using Pólya-gamma augmentation allowed us to explore a variety of model choices without significantly changing our inference procedure: the static LDA, the DTM, and the ccDTM all used the same underlying Gibbs samplers and forward-backward code for Gaussian distributions. It would be interesting to investigate other alternatives, such as correlating intra-document topic proportions with a Pólya-gamma version of the correlated topic model (Blei and Lafferty, 2006a; Linderman et al., 2015) and correlating inter-document topic proportions from the same patient with an author topic model (Rosen-Zvi et al., 2004).

Another interesting direction for exploration is how the same topic appears in different corpora. In our work, we used the simplest approach in which topics for each corpus were isotropically perturbed versions of the latent disease process topic. This approach had the advantage of being able to easily interpret the latent topics probabilities $\boldsymbol{u}_{t,k}$. However, another option might be to learn a static emission matrix $\boldsymbol{C}_l$ for each corpus l:

$$\begin{aligned}\boldsymbol{\beta}_{t,k,l} &\equiv \pi_{\mathsf{SB}}(\boldsymbol{\psi}_{t,k,l}) \\ \boldsymbol{\psi}_{t,k,l} &\equiv \boldsymbol{C}_l \boldsymbol{u}_{t,k} \\ \boldsymbol{u}_{t,k} &\sim \mathcal{N}(\boldsymbol{u}_{t,k} \,|\, \boldsymbol{A}, \boldsymbol{u}_{t-1,k}, \boldsymbol{B}\boldsymbol{B}^{\mathsf{T}})\end{aligned}$$

Such an approach could allow the statistics of the pathological process $\boldsymbol{u}_{t,k}$ to have much lower dimensionality than the corpus-specific topic-word parameters $\boldsymbol{\psi}$ if $\boldsymbol{C}$ were rectangular. It could also model systematic differences between document collections. For example, it would be exciting to incorporate general terms from social media that are not diseases or syndromes. However, the statistics $\boldsymbol{u}_{t,k}$ would be much harder to interpret; we chose our simpler model because $\boldsymbol{u}_{t,k}$ can readily be interpreted as the key terms of the disease process k. To create interpretable reduced-rank models, one approach might be to require that the emission matrix $\boldsymbol{C}_l$ respect some clinician-interpretable ontology, as was done for static topic models in Doshi-Velez et al. (2015).

While Pólya-gamma augmentation allows for the exploration of many exciting models, there are some aspects of the inference that must be treated with care. Our application had a much higher dimensionality than the work of Linderman et al. (2015), and numerical errors accumulated during the recursive stick-breaking construction. Ordering the vocabulary by the prevalence of terms had a large impact on inference performance; deeply understanding the limitations of this augmentation approach on high-dimensionality data sets remains an interesting and open question. Our sampler was also fully uncollapsed; it would be interesting to see whether parameters in the cross-corpora models can be collapsed for

faster-mixing inference. As an alternative inference strategy, black-box variational inference (BBVI) (Ranganath et al., 2014) may offer convenient ways to work with such non-conjugate models.

Clinical Relevance: Autism Spectrum Disorders Clinical manifestations of autism spectrum disorders (ASD) beyond the core DSM criteria have been gaining increasing attention in recent years (Ming et al., 2008; Bauman, 2010; Coury, 2010; Smith, 1981; Kohane et al., 2012). Prior work in clustering phenotypes in ASD has largely relied on surveys and diagnostic tests. Miles et al. (2005) divide ASD into two clusters, "essential" and "complex" based on the manifestation of significant dysmorphology or microcephaly. They find that patients with "complex" ASDs have poorer outcomes, including lower IQ and more seizures. Wiggins et al. (2012) find clusters along disease severity, while Lane et al. (2010) discover sensory processing subtypes. Other studies find clusters along cognitive, language, and behavioral criteria (Wing and Gould, 1979; Ben-Sasson et al., 2008; Bitsika et al., 2008; Hu and Steinberg, 2009). Sacco et al. (2012) find patterns among both neurodevelopmental factors as well as immune and circadian dysfunction.

The phenotypes we find are consistent with these studies as well as the neurological, multi-system, and psychiatric disorder clusters characterized by Doshi-Velez et al. (2013). In addition, we find trajectories for patients with ASD and Down's syndrome and ASD and cerebral palsy, two common comorbidities. Meanwhile, the topics associated with the social media—containing terms such as tantrums and bullying—provide a more complete window in the lives of these children. The fact that mental health terms dominate the social media topics is an indication of important stressors for these children and caregivers.

While it is reassuring that the topics associated with the clinical data are consistent with prior work, this study still has important limitations. Diagnostic codes are extremely noisy measures of disease state, and information extraction from social media is also a challenging process. In particular, our extraction is agnostic to whether a term applies to a current or past condition, to the child or to the caregiver. Our coarse processing was sufficient to discover credible trends, but better extraction methods will be required to validate the patterns we have discussed. Furthermore, while we have shown that our dynamic topic modeling approaches do better at predicting a patient's future diagnoses than static models, there is still an important gap between improved predictions and clinically-useful predictions. Filling this gap will require using additional features in the models and rigorous data validation (e.g. through chart review).

Other Phenotyping Applications While we have focused on developmental disorders, the approaches described here could be relevant to discover the disease trajectories in other conditions. Indeed, almost all disease processes are likely best modeled as continuously evolving rather than having discrete stages. However, applying our approach to complex, chronic diseases such as chronic obstructive pulmonary disease, chronic kidney disease, or diabetes will have several challenges. First, unlike developmental disorders, which start at birth, one must now infer the age of onset from observational sources. Second, while disease processes are continuous, patients often visit when their situation has changed, leading clinicians to observe discrete changes. We hypothesize that a cross-corpora approach, using patient or caregiver-generated text or even outputs of patient-worn sensors (such as glucose monitors), could help discover these continuously evolving processes between sporadic

patient visits. Finally, many of these adult chronic diseases may have periods of remission between periods of high disease activity; these will also need to be modeled.

Conclusions In this work, we presented a dynamic topic modeling approach to modeling disease evolution. Our application of Pólya-gamma augmentation to these models created a simple, unified framework for inference in dynamic topic models and cross-collection topic models. Applied to large collection of EHR and online forum posts describing patients with ASD, our models discovered disease trajectories that make sense in the context of the existing autism literature, and our cross-collection dynamic topic model had both high overall predictive performance and high predictive performance on predicting future patient trajectories. We are excited by the opportunity created by our approach to discover cross-corpora patterns of disease evolution in ASD as well as other diseases.

Acknowledgments

We are grateful to John Bickel and the Boston Children's Hospital i2b2 team for providing the electronic health record data. Joy Ming, Sam Wiseman, and Andy Miller's work on understanding autism forum data was directly valuable for in our pre-processing pipeline. We would also like to acknowledge support for this project from the National Science Foundation (NSF grant ACI-1544628).

References

Ida Anjomshoaa, Margaret E Cooper, and Alexandre R Vieira. Caries is associated with asthma and epilepsy. *European journal of dentistry*, 3(4):297, 2009.

Jon Baio. Prevalence of autism spectrum disorders among children aged 8 years – autism and developmental disabilities monitoring network, 11 sites, united states, 2010. *CDC Morbidity and Mortality Weekly Report*, 63:1–21, March 2014.

M. L. Bauman. Medical comorbidities in autism: challenges to diagnosis and treatment. *Neurotherapeutics*, 7:320327., 2010.

Kevin G Becker. Autism, asthma, inflammation, and the hygiene hypothesis. *Medical hypotheses*, 69(4):731–740, 2007.

David Belanger and Sham Kakade. A linear dynamical system model for text. In *Proceedings of the International Conference on Machine Learning*, 2015.

A. Ben-Sasson, S. A. Cermak, G. I. Orsmond, H. Tager-Flusberg, M. B. Kadlec, and A. S. Carter. Sensory clusters of toddlers with autism spectrum disorders: differences in affective symptoms. *J Child Psychol Psychiatry.*, 49(8):817–25, Aug 2008.

David A Beuther, Scott T Weiss, and E Rand Sutherland. Obesity and asthma. *American Journal of Respiratory and Critical Care Medicine*, 174(2):112–119, 2006.

V. Bitsika, C. F. Sharpley, and S. Orapeleng. An exploratory analysis of the use of cognitive, adaptive and behavioural indices for cluster analysis of asd subgroups. *J Intellect Disabil Res.*, 52(11):973–85, Nov 2008.

David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006a.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the International Conference on Machine Learning*, pages 113–120. ACM, 2006b.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270, 2004.

Coleen A Boyle, Sheree Boulet, Laura A Schieve, Robin A Cohen, Stephen J Blumberg, Marshalyn Yeargin-Allsopp, Susanna Visser, and Michael D Kogan. Trends in the prevalence of developmental disabilities in us children, 1997–2008. *Pediatrics*, pages peds–2010, 2011.

Z. Che, D. Kale, W. Li, M. T. Bahadori, and Y. Liu. Deep computational phenotyping. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.

Brant W Chee, Richard Berlin, and Bruce Schatz. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, volume 2011, page 217. American Medical Informatics Association, 2011.

Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2445–2453, 2013.

Freddy Chong Tat Chua, Richard J Oentaryo, and Ee-Peng Lim. Using linear dynamical topic model for inferring temporal social correlation in latent space. *arXiv preprint arXiv:1501.01270*, 2015.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *NAACL Workshop on Computational Linguistics and Clinical Psychology*, 2015.

D. Coury. Medical treatment of autism spectrum disorders. *Curr Opin Neurol*, 23:131136, 2010.

Willem De Winter, Joost DeJongh, Teun Post, Bart Ploeger, Richard Urquhart, Ian Moules, David Eckland, and Meindert Danhof. A mechanism-based disease progression model for comparison of long-term effects of pioglitazone, metformin and gliclazide on disease processes underlying type 2 diabetes mellitus. *Journal of pharmacokinetics and pharmacodynamics*, 33(3):313–343, 2006.

G. Robert DeLong and Judith T. Dwyer. Correlation of family history with specific autistic subgroups: Asperger's syndrome and bipolar affective disease. *Journal of Autism and Developmental Disorders*, 18:593–600, 1988.

Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*, 10.1542, 2013.

Finale Doshi-Velez, Byron Wallace, and Ryan Adams. Graph-sparse lda: A topic model with structured sparsity. *AAAI*, 2015.

David W Dunn, Joan K Austin, Jaroslaw Harezlak, and Walter T Ambrosius. Adhd and epilepsy in childhood. *Developmental Medicine & Child Neurology*, 45(01):50–54, 2003.

Noémie Elhadad, Shaodian Zhang, Patricia Driscoll, and Samuel Brody. Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 516. American Medical Informatics Association, 2014.

Kai Fan, Marisa Eisenberg, Alison Walsh, Allison Aiello, and Katherine Heller. Hierarchical graph-coupled hmms for heterogeneous personalized health data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.

Takao Furukawa, Kaoru Mori, Kazuma Arino, Kazuhiro Hayashi, and Nobuyuki Shirakawa. Identifying the evolutionary process of emerging technologies: A chronological network analysis of world wide web conference sessions. *Technological Forecasting and Social Change*, 91:280–294, 2015.

P Gail Williams, Lonnie L Sears, and AnnaMary Allard. Sleep problems in children with autism. *Journal of sleep research*, 13(3):265–268, 2004.

Gail T Gillon and Brigid C Moriarty. Childhood apraxia of speech: children at risk for persistent reading and spelling disorder. In *Seminars in speech and language*, volume 28, pages 48–57, 2007.

A Gillott, F Furniss, and A Walter. Anxiety in high-functioning children with autism. *Autism*, 5(3):277–286, September 2001.

Sylvie Goldman, Cuiling Wang, Miran W Salgado, Paul E Greene, Mimi Kim, and Isabelle Rapin. Motor stereotypies in children with autism and other developmental disorders. *Developmental Medicine & Child Neurology*, 51(1):30–38, 2009.

Sudhir Gupta, Daljeet Samra, and Sudhanshu Agrawal. Adaptive and innate immune responses in autism: rationale for therapeutic use of intravenous immunoglobulin. *Journal of Clinical Immunology*, 30(1):90–96, 2010.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Karoly Horvath, John C Papadimitriou, Anna Rabsztyn, Cinthia Drachenberg, and J Tyson Tildon. Gastrointestinal abnormalities in children with autistic disorder. *The Journal of pediatrics*, 135(5):559–563, 1999.

V.W. Hu and M.E. Steinberg. Novel clustering of items from the autism diagnostic interview-revised to define phenotypes within autism spectrum disorders. *Autism Res.*, 2 (2):67–77, Apr 2009.

Kaori Ito, Sima Ahadieh, Brian Corrigan, Jonathan French, Terence Fullerton, Thomas Tensfeldt, Alzheimer's Disease Working Group, et al. Disease progression meta-analysis model in alzheimer's disease. *Alzheimer's & Dementia*, 6(1):39–53, 2010.

Christopher H Jackson, Linda D Sharples, Simon G Thompson, Stephen W Duffy, and Elisabeth Couto. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

Mukund Jha and Noémie Elhadad. Cancer stage prediction based on patient online discourse. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 64–71. Association for Computational Linguistics, 2010.

Lindsey Kent, Joanne Evans, Moli Paul, and Margo Sharp. Comorbidity of autistic spectrum disorders in children with down syndrome. *Developmental Medicine & Child Neurology*, 41(3):153–158, 1999.

I.S. Kohane, A. McMurry, G. Weber, D. MacFadden, and L. Rappaport. The co-morbidity burden of children and young adults with autism spectrum disorders. *PLoS ONE*, 7(4), 2012.

M M Konstantareas and S Homatidis. Ear infections in autistic and normal children. *J Autism Dev Disord*, 17(4):585–94, Dec 1987a.

M Mary Konstantareas and Soula Homatidis. Brief report: Ear infections in autistic and normal children. *Journal of autism and developmental disorders*, 17(4):585–594, 1987b.

A.E. Lane, R.L. Young, A.E. Baker, and M. T. Angley. Sensory processing subtypes in autism: association with adaptive behavior. *J Autism Dev Disord.*, 40(1):112–22, Jan 2010.

Frédéric Laumonnier, Frédérique Bonnet-Brilhault, Marie Gomot, Romuald Blanc, Albert David, Marie-Pierre Moizard, Martine Raynaud, Nathalie Ronce, Eric Lemonnier, Patrick Calvas, et al. X-linked mental retardation and autism are associated with a mutation in the nlgn4 gene, a member of the neuroligin family. *The American Journal of Human Genetics*, 74(3):552–557, 2004.

Li-Ching Lee, Rebecca A Harrington, Brian B Louie, and Craig J Newschaffer. Children with autism: Quality of life and parental concerns. *Journal of autism and developmental disorders*, 38(6):1147–1160, 2008.

Y Li, S Swift, and A Tucker. Modelling and analysing the dynamics of disease progression from cross-sectional studies. *Journal of Biomedical Informatics*, 24(2), 2012.

Scott W. Linderman, Matthew J. Johnson, and Ryan P. Adams. Dependent multinomial models made easy: Stick breaking with the plya-gamma augmentation. In *arXiv:1506.05843*, 2015.

Yu-Ying Liu, Hiroshi Ishikawa, Mei Chen, Gadi Wollstein, Joel S Schuman, and James M Rehg. Longitudinal modeling of glaucoma progression using 2-dimensional continuous-time hidden markov model. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pages 444–451. Springer, 2013.

Lola Luo, Dylan Small, Walter F Stewart, and Jason A Roy. Methods for estimating kidney disease stage transition probabilities using electronic medical records. *EGEMS*, 1 (3), 2013.

Carole L Marcus, Thomas G Keens, Daisy B Bautista, Walter S von Pechmann, and Sally L Davidson Ward. Obstructive sleep apnea in children with down syndrome. *Pediatrics*, 88(1):132–139, 1991.

Tomáš Mikolov. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 2012.

J H Miles, T N Takahashi, S Bagby, P K Sahota, D F Vaslow, C H Wang, R E Hillman, and J E Farmer. Essential versus complex autism: definition of fundamental prognostic subtypes. *Am J Med Genet A.*, 135:171–180, June 2005.

X. Ming, M. Brimacombe, J. Chaaban, B. Zimmerman-Bier, and G. C. Wagner. Autism spectrum disorders: concurrent clinical disorders. *Journal of Child Neurology*, 23:6–13, 2008.

Guillermo Montes and Jill S Halterman. Bullying among children with autism and the influence of comorbidity with adhd: A population-based study. *Ambulatory Pediatrics*, 7 (3):253–257, 2007.

Isager T Mouridsen SE, Rich B. Epilepsy in disintegrative psychosis and infantile autism: a long-term validation study. . *Dev Med Child Neurol*, 41:110114, 1999.

Alistair T Pagnamenta, Elena Bacchelli, Maretha V de Jonge, Ghazala Mirza, Thomas S Scerri, Fiorella Minopoli, Andreas Chiocchetti, Kerstin U Ludwig, Per Hoffmann, Silvia Paracchini, et al. Characterization of a family with rare deletions in cntnap5 and dock4 suggests novel risk loci for autism and dyslexia. *Biological psychiatry*, 68(4):320–328, 2010.

Michael Paul. Cross-collection topic models: Automatically comparing and contrasting text. *Urbana*, 51:61801, 2009.

Michael Paul and Roxana Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods*

in Natural Language Processing: Volume 3-Volume 3, pages 1408–1417. Association for Computational Linguistics, 2009.

Michael Paul, Mark Dredze, David Broniatowski, and Nicholas Generous. Worldwide influenza surveillance through twitter. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*, 2015.

Nienke Peters-Scheffer, Robert Didden, Hubert Korzilius, and Peter Sturmey. A meta-analytic study on the effectiveness of comprehensive aba-based early intervention programs for children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 5(1):60–69, 2011.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association*, 108 (504):1339–1349, 2013.

Teun M Post, Jan I Freijer, Joost DeJongh, and Meindert Danhof. Disease system analysis: basic disease progression models in degenerative disease. *Pharmaceutical research*, 22(7): 1038–1049, 2005.

Siegfried M Pueschel. Clinical aspects of down syndrome from infancy to adulthood. *American Journal of Medical Genetics*, 37(S7):52–56, 1990.

G Ram and J Chinen. Infections and immunodeficiency in down syndrome. *Clinical & Experimental Immunology*, 164(1):9–16, 2011.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

Peder Rasmussen, Ola Börjesson, Elisabet Wentz, and Christopher Gillberg. Autistic disorders in down syndrome: background factors and clinical correlates. *Developmental Medicine & Child Neurology*, 43(11):750–754, 2001.

AL Reyes, AJ Cash, SH Green, and IW Booth. Gastrooesophageal reflux in children with cerebral palsy. *Child: care, health and development*, 19(2):109–118, 1993.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

Ulf Rosenhall, Viviann Nordin, Mikael Sandstrom, Gunilla Ahlsen, and Christopher Gillberg. Autism and hearing loss. *J Autism Dev Disord*, 29(5):349–357, October 1999.

James C. Ross, Peter J. Castaldi, Michael H. Cho, and Jennifer G. Dy. Dual beta process priors for latent cluster discovery in chronic obstructive pulmonary disease. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 155–162, 2014.

R. Sacco, C. Lenti, M. Saccani, C. Paolo, B. Manzi, C. Bravaccio, and A. M. Persic. Cluster analysis of autistic patients based on principal pathogenetic components. *Autism Research*, 5:137–147, 2012.

Ardavan Saeedi and Alexandre Bouchard-Côté. Priors over recurrent continuous time processes. In *Advances in Neural Information Processing Systems*, pages 2052–2060, 2011.

Michael B Shapiro and Thomas D France. The ocular features of down's syndrome. *American journal of ophthalmology*, 99(6):659–663, 1985.

Andrew J Sharp, Heather C Mefford, Kelly Li, Carl Baker, Cindy Skinner, Roger E Stevenson, Richard J Schroer, Francesca Novara, Manuela De Gregori, Roberto Ciccone, et al. A recurrent 15q13. 3 microdeletion syndrome associated with mental retardation and seizures. *Nature genetics*, 40(3):322–328, 2008.

E H Sherr. The arx story (epilepsy, mental retardation, autism, and cerebral malformations): one gene leads to many phenotypes. *Curr Opin Pediatr*, 6(15):567–571, December 2003a.

Elliott H Sherr. The arx story (epilepsy, mental retardation, autism, and cerebral malformations): one gene leads to many phenotypes. *Current opinion in pediatrics*, 15(6): 567–571, 2003b.

Sally R Shott, Aileen Joseph, and Dorsey Heithaus. Hearing loss in children with down syndrome. *International journal of pediatric otorhinolaryngology*, 61(3):199–205, 2001.

R.D. Smith. Abnormal head circumference in learning-disabled children. *Dev Med Child Neurol*, 23:626632, 1981.

Alexander E Sochaniwskyj, Ruth M Koheil, Kazek Bablich, Morris Milner, and David J Kenny. Oral motor functioning, frequency of swallowing and drooling in normal children and in children with cerebral palsy. *Archives of physical medicine and rehabilitation*, 67 (12):866–874, 1986.

Rich Stoner, Maggie L Chow, Maureen P Boyle, Susan M Sunkin, Peter R Mouton, Subhojit Roy, Anthony Wynshaw-Boris, Sophia A Colamarino, Ed S Lein, and Eric Courchesne. Patches of disorganization in the neocortex of children with autism. *New England Journal of Medicine*, 370(13):1209–1219, 2014.

Rafid Sukkar, Edward Katz, Yanwei Zhang, David Raunig, and Bradley T Wyman. Disease progression modeling using hidden markov models. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2845–2848. IEEE, 2012.

Pål Surén, Inger Johanne Bakken, Heidi Aase, Richard Chin, Nina Gunnes, Kari Kveim Lie, Per Magnus, Ted Reichborn-Kjennerud, Synnve Schjølberg, Anne-Siri Øyen, et al. Autism spectrum disorder, adhd, epilepsy, and cerebral palsy in norwegian children. *Pediatrics*, 130(1):e152–e158, 2012.

Michael E. Talkowski, Gilles Maussion, Liam Crapper, Jill A. Rosenfeld, Ian Blumenthal, Carrie Hanscom, Colby Chiang, Amelia Lindgren, Shahrin Pereira, Douglas Ruderfer, Alpha B. Diallo, Juan Pablo Lopez, Gustavo Turecki, Elizabeth S. Chen, Carolina Gigek, David J. Harris, Va Lip, Yu An, Marta Biagioli, Marcy E. MacDonald, Michael Lin, Stephen J. Haggarty, Pamela Sklar, Shaun Purcell, Manolis Kellis, Stuart Schwartz, Lisa G. Shaffer, Marvin R. Natowicz, Yiping Shen, Cynthia C. Morton, James F. Gusella, and Carl Ernst. Disruption of a large intergenic noncoding rna in subjects with neurodevelopmental disabilities. *The American Journal of Human Genetics*, 91:1128–1134, December 2012.

Stephen W Thomas, Bram Adams, Ahmed E Hassan, and Dorothea Blostein. Studying software evolution using topic models. *Science of Computer Programming*, 80:457–479, 2014.

Eeske Van Roekel, Ron HJ Scholte, and Robert Didden. Bullying among adolescents with autism spectrum disorders: Prevalence and perception. *Journal of autism and developmental disorders*, 40(1):63–73, 2010.

Chong Wang, Bo Thiesson, Christopher Meek, and David Blei. Markov topic models. In *Proceedings of The Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.

Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*, 2012.

Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

Megan L Wier, Roxana Odouli, Cathleen K Yoshida, Judith K Grether, and Lisa A Croen. Congenital anomalies associated with autism spectrum disorders. *Developmental Medicine & Child Neurology*, 48(6):500–507, 2006.

L. D. Wiggins, D. L. Robins, L. B. Adamson, R. Bakeman, and C. C. Henrich. Support for a dimensional view of autism spectrum disorders in toddlers. *J Autism Dev Disord.*, 42 (2):191–200, Feb 2012.

Kumanan Wilson and John S Brownstein. Early detection of disease outbreaks using the internet. *Canadian Medical Association Journal*, 180(8):829–831, 2009.

Jesse Windle, Nicholas G Polson, and James G Scott. Sampling Pólya-gamma random variates: alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*, 2014.

L. Wing and J. Gould. Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification. *J Autism Dev Disord.*, 9(1):11–29, Mar 1979.

Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd international conference on World wide web*, pages 783–794. ACM, 2014.

Qing Treitler Zeng. Consumer health vocabulary initiative, 2015. URL `http://consumerhealthvocab.org/`.

ChengXiang Zhai, Atulya Velivelli, and Bei Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 743–748. ACM, 2004.

Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1088. ACM, 2010.

Jiayu Zhou, Jun Liu, Vaibhav A Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103. ACM, 2012a.

Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 135–144. ACM, 2014.

Mingyuan Zhou, Lingbo Li, David Dunson, and Lawrence Carin. Lognormal and gamma mixed negative binomial regression. In *Proceedings of the International Conference on Machine Learning*, volume 2012, page 1343, 2012b.

Appendix A. Data and Data Processing

A.1 Example Forum Post

Below is an example of a post. The age and CUIs that we extracted from the post are listed below.

```
hi my son is 13 nearly 14 and has this year become increasingly
 anxious and withdrawn in july his psychiatrist said to put him on
 prozac saying it might take the edge off his anxieties and allow him
 some positive experiences thus helping to lift the depression he
 seemed to be in i was not all that keen to be honest but my son who
 had been reluctant to take his other meds said he wanted to try it so
 we did he started on a small liquid dose and is now on tab a day will
 check exact dose if you want to know it despite my reservations his
 mood has really lifted he is still really challenging aggressive one
 track mind struggles to leave the house though maybe not so much but
```

```
 to be honest he is back to where he was before the dip in terms of
 talking to me etc i am probably not explaining very well in feb half
 term adn easter hols the only interaction at all was to be negative
 call us names adn swear at us now he still does that but he also
 chats and has a laugh again which had stopped i have not seen any
 side effects and he says he likes taking it cos he feels better he
 cant explain anymore than that i discussed it with autism outreach
 recently and she said it is being used effectively in a lot of kids
 with asd and anxieties to take the edge off the anxieties dont get me
 wrong it hasn t solved all our issues at all but he just doesnt seem
 so saddont know if this is of any help at all so hard to put into
 words lol ps if you google most meds for kids ritalin prozac
 respiridone etc you get a lot of negatives adn not many positives and
 not a lot of balanced comment

age: 13

CUI: C0870663, C0424092, C0683607, C0023133, C0234856, C1273517,
 C1304698, C0001807, C0080151, C0011570, C0233730, C0004352
```

A.2 Forum Data Pre-Processing and Age Extraction

We used BeautifulSoup to obtain and parse all subforums of the websites `www.asd-forum.org.uk`, `www.autismweb.com`, and `www.asdfriendly.org` on June 29, 2015. We extracted the text, the user-id, and the time and date of posting for 21,206 threads from asd-forum, 26,807 threads from asd-friendly, and 32,914 threads from autismweb, for a total of 80,927 threads. These threads contained a total of 664,954 posts. Figure 8 shows the regular expressions used to extract ages from the posts. Next, the outputs were filtered through a trie for a list of error terms such as sec, wks, ft, and m that might indicate another unit of measure; posts with such terms after the identified age were excluded. Finally, only posts with only one age were included, to avoid conflating information from multiple ages or multiple people.

Appendix B. MCMC Convergence

Figure 9 show the log-likelihoods for a characteristic run. Based on plots such as these, we determined that 100 iterations seemed to be more than enough for the sampler to find an optima.

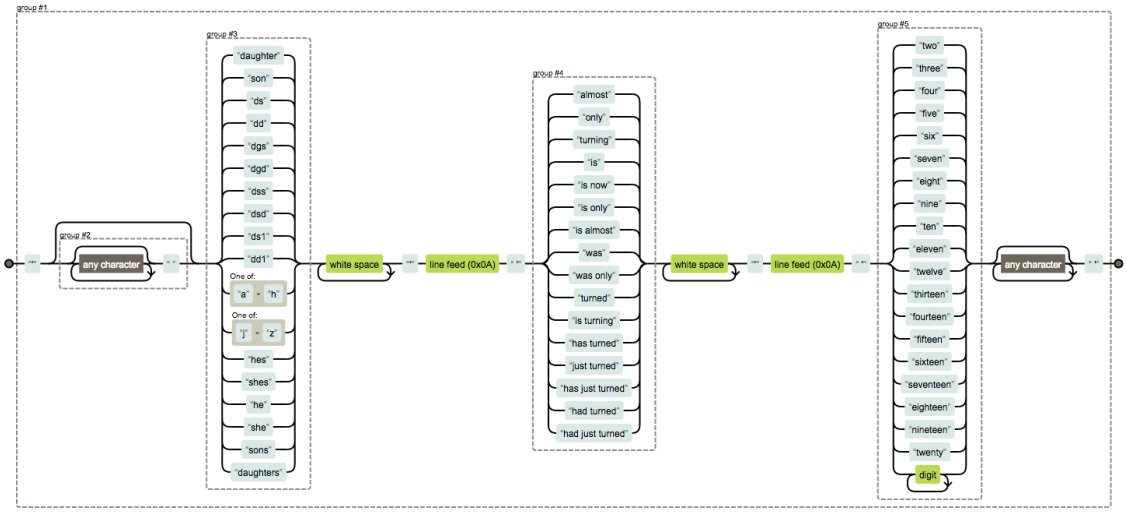


Figure 8: Chart showing the of regular expressions used to extract potential ages from posts. Outputs passing this filter were filtered through a second stage of processing to identify and remove cases where the number corresponded to a unit other than age in years.

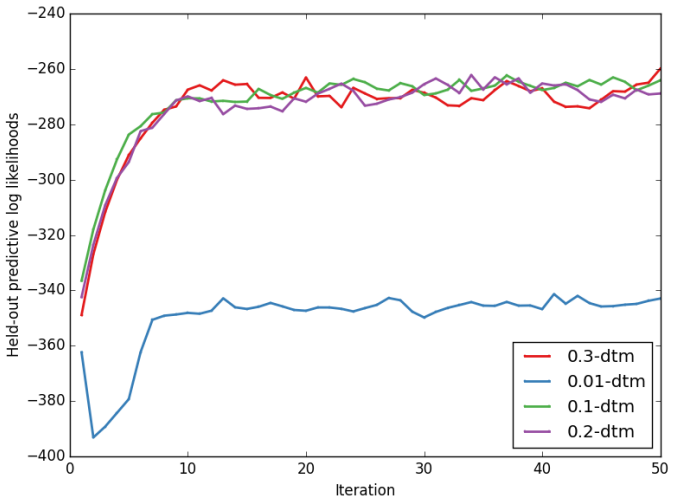


Figure 9: Log-likelihoods for a characteristic run.