Incorporating External Information in Tissue Subtyping: A Topic Modeling Approach

Ardavan Saeedi[*][†] AV.SAEEDI@GMAIL.COM
Hyperfine

Payman Yadollahpour[*] PYADOLLA@BROADINSTITUTE.ORG
Broad Institute

Sumedha Singla SUMEDHA.SINGLA@PITT.EDU
University of Pittsburgh

Brian Pollack BRP98@PITT.EDU
University of Pittsburgh

William Wells SW@BWH.HARVARD.EDU
Harvard Medical School / Brigham and Women's Hospital

Frank Sciurba CIURBAFC@UPMC.EDU
University of Pittsburgh Medical Center

Kayhan Batmanghelich KAYHAN@PITT.EDU
University of Pittsburgh

Abstract

Probabilistic topic models, have been widely deployed for various applications such as learning disease or tissue subtypes. Yet, learning the parameters of such models is usually an ill-posed problem and may result in losing valuable information about disease severity. A common approach is to add a discriminative loss term to the generative model's loss in order to learn a representation that is also predictive of disease severity. However, finding a balance between these two losses is not straightforward. We propose an alternative way in this paper. We develop a framework which allows for incorporating external covariates into the generative model's approximate posterior. These covariates can have more discriminative power for disease severity compared to the representation that we extract from the posterior distribution. For instance, they can be features extracted from a neural network which predicts disease severity from CT images. Effectively, we enforce the generative model's approximate posterior to reside in the subspace of these discriminative covariates. We illustrate our method's application on a large-scale lung CT study of Chronic Obstructive Pulmonary Disease (COPD), a highly heterogeneous disease. We aim at identifying tissue subtypes by using a variant of topic model as a generative model. We quantitatively evaluate the predictive performance of the inferred subtypes and demonstrate that our method outperforms or performs on par with some reasonable baselines. We also show that some of the discovered subtypes are correlated with genetic measurements, suggesting that the identified subtypes may characterize the disease's underlying etiology.

[*] Equal contribution

[†] Work is not related to the research done at Hyperfine.

1. Introduction

Probabilistic models have been widely used to uncover hidden phenotypes for various healthcare applications, such as inferring rates of aging (Pierson et al., 2019), survival prediction (Chen and Weiss, 2017), disease subtyping (Batmanghelich et al., 2015), and many more (Chen et al., 2020). One of the challenges of applying the generative models in medical applications is to ensure that the inferred parameters reflect the disease status; for example, the proportion of abnormal tissue subtype in each patient should be correlated with the clinical measurements reflecting the disease severity. We develop a model that allows for incorporating external covariates into the posterior inference. The external covariates can be flexibly designed such that they are correlated with the disease severity. For instance, these covariates can be features extracted from a neural network predicting clinical measurements.

We apply our approach in the context of Chronic Obstructive Pulmonary Disease (COPD), which is a highly heterogeneous disease (Castaldi et al., 2017b; Chen et al., 2013). COPD is characterized by inflammation of the airway and destruction of the air sacs (emphysema) (Viegi et al., 2007), and is one of the leading causes of death worldwide (Decramer et al., 2012; World Health Organization, 2018). There are differences between risk factors of COPD subtypes (Shapiro, 2000), and hence understanding subtypes is important. Respirometry measurement is used for the diagnosis of COPD; however, it cannot identify the underlying process of COPD. Hence, computed tomography (CT) imaging, which allows direct qualitative and quantitative evaluation of tissue destruction, is routinely requested for COPD patients. For example, phenotypic abnormality of emphysema is evident from CT images (Park et al., 2008; Ross et al., 2016). Although there has been significant work on defining *visual* subtypes of emphysema (Song et al., 2017; Ross et al., 2016; Yang et al., 2017; Häme et al., 2015; Uppaluri et al., 1997; Sorensen et al., 2010; Depeursinge et al., 2007; Prasad et al., 2009) from CT images, there is significant intra-reader and inter-reader variability of visual subtypes (Binder et al., 2016; Aziz et al., 2004). In this paper, we adopt a variant of topic modeling to formulate the subtype discovery problem.

We view the CT image of every patient as a mixture of K typical imaging patterns that reoccur across the population. The proportion of the mixture is patient specific, but the patterns are shared across the population. We call the typical pattern “tissue subtype.” This way of explaining data is reminiscent of topic models where the topics are tissue subtypes. Hence, we use “subtype” and “topic” interchangeably. The distribution of each patient's tissue subtype can be viewed as the patient representation. Off-the-shelf topic modeling is unsupervised, and it focuses on explaining the data and can easily miss the disease-relevant information. We aim to address this issue in this paper. We enforce the patient representation to be correlated with disease severity, and hence indirectly encourage subtypes to be disease-related. Instead of supervised topic modeling, we propose to incorporate discriminative information in the form of covariates into the subtypes' inference model (*i.e.,* topics).

Related Works. Various unsupervised phenotype discovery methods have been proposed in the healthcare domain (*e.g.,* Pivovarov et al. (2015); Urteaga et al. (2020)). Image-based phenotype discovery in CT images via spatial texture patterns have been explored in emphysema (Yang et al., 2017; Häme et al., 2015). Ross et al. (2016) propose a generative graphical model that incorporates patient trajectories to identify disease subtypes for COPD. Binder et al. (2016) present a generative model for unsupervised discovery of visual subtypes for COPD along with inferring population structure. Their method identifies subpopulations and clusters of image pattern simultaneously. One of the underlying assumptions of these methods is that the patient population can be divided into sub-populations, which is disputed for COPD (Castaldi et al., 2017a). Furthermore, these methods are unsupervised – solving a highly ill-posed problem – hence, the resulting subtypes may not reflect disease severity.

On the other hand, many supervised methods have been proposed to characterize the severity of lung diseases from CT images (Uppaluri et al., 1997; Depeursinge et al., 2007; Park et al., 2008; Prasad et al., 2009; Sorensen et al., 2010; Walsh et al., 2018). These methods study local descriptors such as local binary pattern (LBP) (Sorensen et al., 2010), wavelet and gray-level features (Depeursinge et al., 2007) as well as various predictive methods ranging from $k-$nearest neighbor classifier (Sorensen et al., 2010) to Support Vector Machine (SVM) (Park et al., 2008). However, it is not clear how these methods can inform subtype discovery.

Our model is closely related to supervised topic models (Mcauliffe and Blei, 2008; Korshunova et al., 2019; Ren et al., 2019; Lacoste-Julien et al., 2009; Ramage et al., 2009; Hughes et al., 2018) which generally add a discrminative loss term and predict the labels from the topics or topic proportions. In healthcare applications other than COPD, Yang *et al.* (Yang et al., 2019) proposed a supervised topic modeling to characterize Alzheimer's disease subtypes.

Our proposed approach is different from the previous works in three ways:

1. Rather than modeling the disease cohort into sub-populations, we view it as a continuum where the continuum represents the proportion of subtypes. We aim at discovering subtypes across the disease cohort; each patient is a mixture of these subtypes which we assume are manifested in the CT images. The image signature of the subtypes and the patient-specific mixture are modeled as latent variables in a probabilistic generative model and, more specifically, a *topic model* (Blei et al., 2003).

2. We assume that discriminative covariates are provided as extra information. We construct such covariates based on a generic approach and without making any parametric assumption over the model or probability distribution.

3. Unlike supervised topic modeling, our model does not require balancing the generative and discriminative losses; hence, it has fewer hyper-parameters. We propose to incorporate the discriminative covariates into the approximate posterior distribution.

We apply our method on a large scale COPD study showing good predictive performance and clinically interpretable subtypes. Three of the subtypes are shown to have significant genetic heritability. Furthermore, we compare our model with variants of topics models and demonstrate that it outperforms them in terms of predictive performance.

Generalizable Insights about Machine Learning in the Context of Healthcare

This paper makes the following contributions which are generalizable to other applications in healthcare:

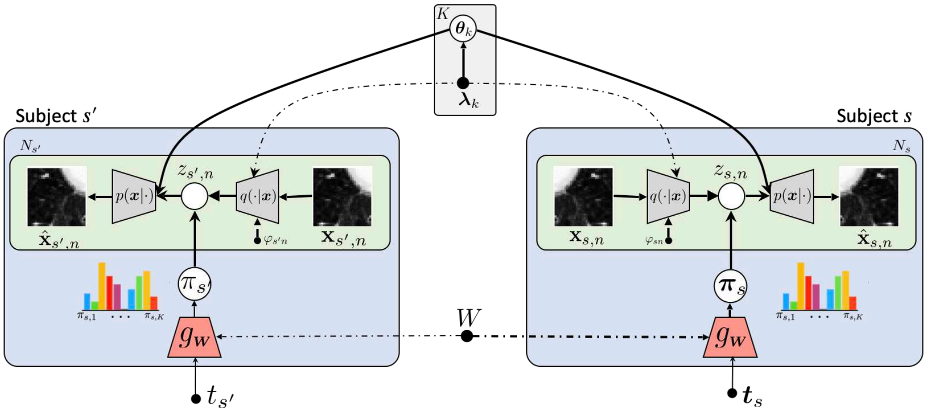


Figure 1: The schematic of our framework for two subjects s and s' with t_s and $t_{s'}$ as their corresponding covariates. The encoder $(q(\cdot|x))$ and decoder $(p(x|\cdot))$ inside the green box explain data at the supervoxel-level (word-level) while g_w explains the subject-level data (*i.e.,* topic proportion). θ_k and λ_k are the parameter of the likelihood function and its corresponding variational parameter. The dashed line denotes sharing the parameters. See Table 1 for the definitions of notations used in this paper.

- We develop a framework for generative disease subtyping that allows for incorporating external covariates into the posterior distribution approximation. We propose an efficient formulation for the posterior approximation that does not incur the extra computational cost during inference and does not require a hyper-parameter to balance supervised and unsupervised loss terms (as in supervised topic models). Although our framework demonstrates promising results on topic models, it can be applied to other probabilistic graphical models that benefit from supervision (*e.g.,* latent factor analysis (Farouni, 2017), mixture models (Hannah et al., 2011) or hidden Markov models for predicting length of stay in ICU (Sotoodeh and Ho, 2019)).

- We apply our framework to disease subtyping based on CT images; however, its use case is not limited to this data type and can be applied to any data type in healthcare for which topic models have shown to be useful. Examples include, topic model application to Electronic Health Records (EHR) (Li et al., 2020), transcriptomic data (Valle et al., 2020), and histopathology data (Cruz-Roa et al., 2011).

- We use covariates that are predictive of disease severity; however, our framework is naturally capable of incorporating other types of relevant side information such as clinical, genetic, and demographic covariates.

2. Method

To represent each subject, we adopt the Bag of Words (BOW) model (Fei-Fei and Perona, 2005) and represent a subject s with a *set*, $\mathcal{X}_s$, containing features extracted from N_s regions covering the lung regions of the subject. This modeling choice allows us to accommodate lungs of different sizes; the number of elements in $\mathcal{X}_s$ can vary depending on the size of the lungs. The BOW model assumes that features of every subject, $\mathbf{x}_{sn} \in \mathcal{X}_s$, are drawn

Decoder	
S	Total number of subjects.
K	Total number of subtypes.
N_s	Number of supervoxels in subject s.
$\boldsymbol{x}_{s,n}$	Image descriptor of supervoxel n in subject s.
$\mathcal{X}_s$	Set of all image features for subject s, ($\boldsymbol{x}_{s,n} \in \mathcal{X}_s$).
$z_{s,n}$	Subject-specific subtype that generates super-voxel n in subject s.
$\boldsymbol{\pi}_s$	Proportions of subtypes in subject s.
$\boldsymbol{\theta}_k$	Parameters of the likelihood (*e.g.,* mean $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ covariance matrix) of image descriptors for population-level subtype k.
$\boldsymbol{\beta}$	Stick-breaking proportions for the Dirichlet Process which defines $\boldsymbol{\pi}_s$.
α	Concentration parameters of the stick-breaking distribution for $\boldsymbol{\beta}$.
Encoder	
$\boldsymbol{\varphi_{s,n}}$	Parameters of the variational posterior for $z_{s,n}$
$\boldsymbol{\omega}_s$	Parameters of the variational posterior for $\boldsymbol{\pi}_s$.
$\boldsymbol{\lambda}_k$	Parameters of the variational posterior for $\boldsymbol{\theta}_k$.
β^*	Parameters encoding the posterior distribution of $\boldsymbol{\beta}$.
$\boldsymbol{t}_s$	Subject-level feature vector.
$\boldsymbol{W}$	Parameters encoding the posterior topic proportions $\boldsymbol{\pi}_s$.
$h_{SB}(\cdot)$	Stick-breaking function.
$\boldsymbol{\psi}_s$	Unnormalized subject-level topic proportions.

Table 1: Summary of the notation used for the decoder (*i.e.,* generative model) and encoder (*i.e.,* variational Bayes posterior approximation) in our proposed framework.

from subject-specific probability distributions, *i.e.,* $\mathbf{x}_{sn} \sim p_s$. We assume that p_s belongs to some abstract space of distributions (*i.e.,* $p_s \in \mathcal{P}$). Our model can be viewed as an encoder-decoder, where the decoder formulates the topic model, and the approximate posterior distribution is formulated by the encoder. Our goal is to approximate the topics' posterior distribution and not image reconstruction. Therefore, to explain features of each topic, we use a parametric model with limited complexity whose expectations, entropy and marginal can be computed efficiently.

In Sections 2.1 and 2.2, we explain our design for the decoder as well as the encoder allowing arbitrary covariate information to be incorporated into inference. The schematic of the framework is given in Fig. 1.

2.1. Decoder

We first explain the probabilistic graphical model that defines the decoder (*i.e.,* generative model). Our model is based on topic modeling, where the topic parameters correspond to the population-level parameters, and document-specific topic proportions correspond to the subject-level distribution of subtypes. In the following, we discuss the modeling assumptions in detail.

Population-Level Model The model assumes that there are K tissue types, *topics*, that are shared across subjects in the population. We use a D-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_k \in \mathbb{R}^D$ and covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^D \times \mathbb{R}^D$ to model the features of the topic k. For computational reasons, we also assume a conjugate prior for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$,

$$\boldsymbol{\theta}_k := (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim \mathrm{NIW}(\eta),$$

where $\mathrm{NIW}(\eta)$ is the Normal-Inverse-Wishart distribution with hyper-parameter η. Note that $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ are random variables not parameters; hence, we aim at estimating a posterior distribution not a point estimate. For notational brevity, let $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Subject-Level Model For subject s, $\boldsymbol{\pi}_s = [\pi_{s1}, \cdots, \pi_{sK}]$ and $\{z_{sn}\}_{n=1}^{N_s}$ are latent random variables denoting the proportion of topics and the allocation of the supervoxels to the topics (*i.e.*, $z_{sn} \in [1 \cdots K]$) respectively:

$$\begin{aligned} \boldsymbol{\pi}_s | \boldsymbol{\beta} &\sim \mathrm{Dir}(\beta_1, \cdots, \beta_K), \\ z_{sn} | \boldsymbol{\pi}_s &\sim \mathrm{Cat}(\boldsymbol{\pi}_s), \\ \mathbf{x}_{sn} | z_{sn}, \{\boldsymbol{\theta}_k\}_{k=1}^K &\sim \mathcal{N}\left(\mu_{z_{sn}}, \Sigma_{z_{sn}}\right); \end{aligned} \tag{1}$$

where the $\boldsymbol{\pi}_s$ follows the Dirichlet distribution, $\mathrm{Cat}(\boldsymbol{\pi}_s)$ represents a categorical distribution with the topic proportion $\boldsymbol{\pi}_s$, and $z_{sn} = k$ indicates supervoxel n of subject s follows the local image descriptor of topic k. The β_k's are concentration parameters. If β_k's are greater than one, the topics distribution becomes more disperse (less sparse).

To avoid tuning K hyper-parameters for β_1 to β_K, we follow the truncated Hierarchical Dirichlet Process (HDP) (Teh et al., 2006), and assume β is generated by the "stick-breaking" construction,

$$\begin{aligned} \tau_j &\overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \alpha), \\ \beta_k &:= \tau_k \prod_{j<k} (1 - \tau_j), \end{aligned} \tag{2}$$

where $\mathrm{Beta}(\cdot, \cdot)$ indicates the Beta distribution. Such construction allows for controlling the sparseness of the topics distribution with a single hyper-parameter (*i.e.*, α) rather than K. Similar to the approach introduced by Bryant and Sudderth (2012), we choose a large enough K and allow the actual number of topics to be discovered from data.

Overall Decoder Model For notational convenience, we define $\mathcal{D} = \{\mathcal{X}_s\}_{s=1}^S$ to be all image data, $\mathcal{S} = \{z_{sn}, \boldsymbol{\pi}_s\}_{s=1}^S$ to be all subject-level latent variables, and $\mathcal{C} = \{\boldsymbol{\theta}_k, \boldsymbol{\beta}\}$ to be all population-level latent variables. The joint distribution of all random variables can be written as follows,

$$p(\mathcal{D}, \mathcal{S}, \mathcal{C}) = p(\boldsymbol{\beta}|\alpha) \prod_k p(\boldsymbol{\theta}_k|\eta) \prod_s p(\boldsymbol{\pi}_s|\boldsymbol{\beta}) \prod_{s,n} p(\boldsymbol{x}_{sn}|z_{sn}, \{\boldsymbol{\theta}_k\}) p(z_{sn}|\boldsymbol{\pi}_s).$$

2.2. Encoder

We propose to incorporate external covariates into the estimation of the posterior distribution. If the covariates are highly correlated with the disease severity, the inferred subtypes will respect the discriminative signal about the disease severity. Our proposed approach is general and can incorporate any external covariate depending on the application. We use $\boldsymbol{t}_s$ to denote the covariate features. First, we explain the classical approach, and then explain our method to incorporate $\boldsymbol{t}_s$.

Variational Bayes (VB) Approximate of the Posterior We seek the true posterior distribution of the model parameters,

$$p(\mathcal{S}, \mathcal{C}|\mathcal{D}) = \frac{p(\mathcal{D}, \mathcal{S}, \mathcal{C})}{\int p(\mathcal{D}, \mathcal{S}, \mathcal{C}) d\mathcal{S} d\mathcal{C}}. \tag{3}$$

Exact computation of the posterior is computationally intractable since the denominator is hard to compute. Therefore, Variational Bayes (M. Blei et al., 2016; Jordan et al., 1999) approximates the posterior by maximizing the Evidence Lower Bound (ELBO) with respect to q,

$$\max_{q \in \mathcal{Q}} \mathcal{L}(q), \quad \mathcal{L}(q) \triangleq \mathbb{E}_q \left[\ln p(\mathcal{D}, \mathcal{S}, \mathcal{C})\right] - \mathbb{E}_q \left[\ln q(\mathcal{S}, \mathcal{C})\right], \tag{4}$$

where $q \in \mathcal{Q}$ is an approximate distribution from the family of computationally efficient probability densities $\mathcal{Q}$. As it is common in mean-field variational inference (Peterson and Anderson, 1987; Jordan et al., 1999; Hoffman et al., 2013; M. Blei et al., 2016), we assume the following form for the approximate posterior, $q(\cdot)$,

$$\mathcal{Q} : q(\mathcal{S}, \mathcal{C}) = q(\boldsymbol{\beta}; \beta^*) \underbrace{\prod_s q(\boldsymbol{\pi}_s; \boldsymbol{\omega}_s)}_{\text{subject-level}} \underbrace{\prod_{s,n} q(z_{sn}; \boldsymbol{\varphi}_{sn})}_{\text{spatial level}} \underbrace{\prod_k q(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k)}_{\text{population-level}}, \tag{5}$$

where β^*, $\boldsymbol{\varphi}_{sn}$, $\boldsymbol{\lambda}_k$, and $\boldsymbol{\omega}_s$ are the variational parameters corresponding to the random variables β, z_{sn}, $\boldsymbol{\theta}_k$, and $\boldsymbol{\pi}_s$, respectively.

We use the variational parameters of $q(\mathcal{S}, \mathcal{C})$ to approximate the posterior distribution of the *population-level*, *subject-level*, and *spatial level* variables. Specifically, we approximate (1) the posterior distribution of $\boldsymbol{\theta}_k$'s as the image descriptors of each subtype (topic), (2) the posterior distribution of $\boldsymbol{\pi}_s$ as the proportion of subtypes per subject and (3) the posterior distribution of $z_{s,.}$ that visualizes the spatial distribution of the subtypes within the lung of patient s. The exact parametric form for each term is given in Appendix C.

Incorporating the Covariates into Posterior Approximation In the previous sections, we described the standard topic model construction and the corresponding family of variational distributions used to approximate the posterior of the latent variables in the model. The standard inference method for topic modeling does not allow for incorporating the external covariates. We define a new family of approximate posterior distributions, $\mathcal{Q}'$, that allows for the external covariates without incurring an extra computational cost during inference[1].

Unlike the rest of the variables, $\boldsymbol{\pi}_s$ is defined at the *subject-level,* characterizing the topic proportions for subject s. We also have $\boldsymbol{t}_s$ which is a subject-specific covariate. Hence, we introduce $\boldsymbol{t}_s$ to the posterior of the $\boldsymbol{\pi}_s$. To do that, we use $\boldsymbol{t}_s$, the subject-specific representation, to encode the subject-level latent variable. In other words, we use $\boldsymbol{t}_s$ to parameterize the variational posterior for $\boldsymbol{\pi}_s$: $q(\boldsymbol{\pi}_s|\boldsymbol{t}_s; \boldsymbol{W})$, where $\boldsymbol{W} = \{\boldsymbol{W}_{\boldsymbol{\sigma}}, \boldsymbol{W}_{\boldsymbol{\mu}}\}$ is a new parametrization of the latent variables $\boldsymbol{\pi}_s$. Note that previously we had different variational

[1] Note that, depending on the covariates, there might be extra computational costs (*e.g.,* cost of training a neural network) for obtaining the covariates.

parameters $\boldsymbol{\omega}_s$ for each subject; we now have one set of parameters $\boldsymbol{W}$ shared across all subjects.

We model $q(\boldsymbol{\pi}_s)$ implicitly by sampling from a Gaussian distribution and passing the samples through a function to normalize them to a simplex (*i.e.*, $\sum_k [\boldsymbol{\pi}_s]_k = 1$). Similar to the idea of reparameterization trick in Variational Autoencoder (VAE) (Kingma and Welling, 2013), we parameterize the mean and variance of the Gaussian by a neural network. However, instead of inputting the original image, we use the subject-level representation, $\boldsymbol{t}_s$, as input:

$$\begin{aligned} \boldsymbol{\epsilon} &\sim \mathcal{N}(0, I_{K\times K}) \\ \boldsymbol{\psi}_s &= \boldsymbol{\mu}(\boldsymbol{t}_s; \boldsymbol{W}_{\boldsymbol{\mu}}) + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}(\boldsymbol{t}_s; \boldsymbol{W}_{\boldsymbol{\sigma}}) \\ \boldsymbol{\pi}_s &= h_{SB}(\boldsymbol{\psi}_s), \end{aligned} \tag{6}$$

where $\boldsymbol{\mu}(\boldsymbol{t}_s; \boldsymbol{W}_{\boldsymbol{\mu}})$ and $\boldsymbol{\sigma}(\boldsymbol{t}_s; \boldsymbol{W}_{\boldsymbol{\sigma}})$ are neural networks computing the mean and variance vector of $\boldsymbol{\psi}_s$, respectively. The $h_{SB}(\cdot)$ is a function transforming the unbounded values of $\boldsymbol{\psi}_s$ drawn from a Gaussian distribution to a random variable on a simplex, *i.e.*, $h_{SB} : \mathbb{R}^K \rightarrow \Delta^K$. Many choices are possible for $h_{SB}(\cdot)$, such as the *softmax* function. However, computing the probability density of the transformed random variable is not always straightforward. Here, we choose the following form that enables us to have a closed-form probability density for $\boldsymbol{\pi}_s$ (Linderman et al., 2015),

$$h_{SB}(\boldsymbol{\psi}_s): \quad \boldsymbol{\pi}_{sk} \; = \sigma(\boldsymbol{\psi}_{sk})(1 - \sum_{j<k} \boldsymbol{\pi}_{sj}), \tag{7}$$

where $\sigma(\cdot)$ denotes the logistic function. The $\boldsymbol{\pi}_s$, which is the result of a change of variable, has the following probability density,

$$q(\boldsymbol{\pi}_s | \boldsymbol{t}_s; \boldsymbol{W}) = \mathcal{N}(\boldsymbol{\psi}_s; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)) \left| \left\{ \frac{\partial [\boldsymbol{\pi}_s]_i}{\partial [\boldsymbol{\psi}_s]_j} \right\} \right|^{-1}, \tag{8}$$

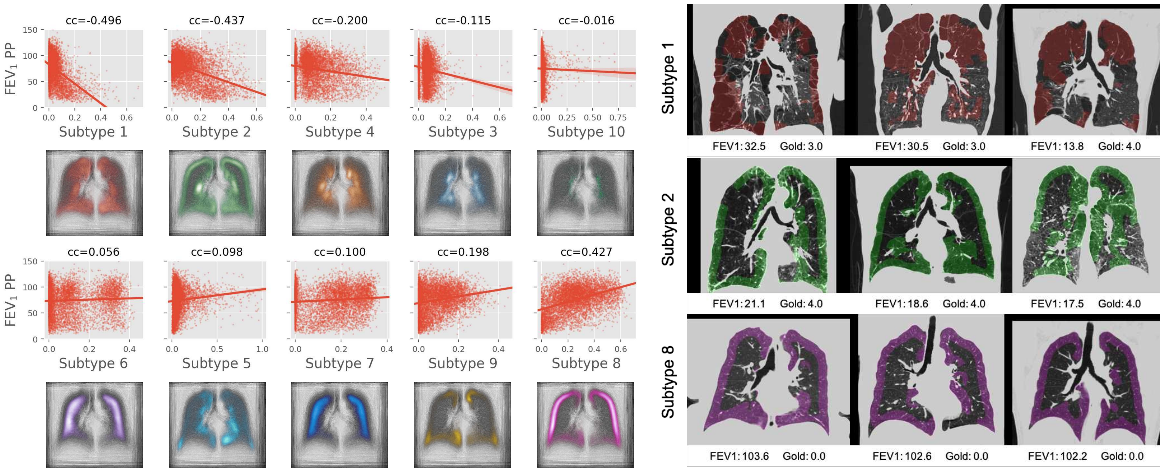
where $\left| \left\{ \frac{\partial [\boldsymbol{\pi}_s]_i}{\partial [\boldsymbol{\psi}_s]_j} \right\} \right|$ is the determinant of the Jacobian which is easily computable (see Appendix C). This is a computationally appealing property for our optimization-based inference as we can easily plug it into the factorization of $q(\mathcal{S}, \mathcal{C})$.

Similar to the classical model in Section 2.2, the parameters of this model are learned by maximizing the ELBO. All updates have a similar form as before except $\boldsymbol{W}_{\boldsymbol{\mu}}$ and $\boldsymbol{W}_{\boldsymbol{\sigma}}$, for which we use stochastic gradient descent (see Appendix C for more details).

3. Experiments

In this section, we evaluate the proposed method for lung tissue subtyping on a large-scale dataset from the COPDGene study (Regan et al., 2011). In Section 3.1, first we describe the dataset we use for evaluation. Next, we explain our feature extraction pipeline and the clinical measurements that we use for evaluation.

In Section 3.2, we demonstrate that the extracted features are informative by comparing them with a set of reasonable baselines in terms of being able to predict the clinical measurements. Next we compare the predictive performance of our framework, with that of a topic model and a supervised variant of it.



(a) (b)

Figure 2: (a) *Odd Rows:* Pearson correlation between proportion of subtype and FEV_1. The x- and y-axis are the subtype proportion and FEV_1 respectively. *Even Rows:* Visualization of spatial average of the learned subtypes across the population shown on a coronal slice of a lung atlas. (b) *Subtypes 1*, *2*, and *8* depicted on a set of nine patients. *Subtypes 1* and *2* are correlated with increase in severity of COPD (negatively correlated with FEV_1), whereas *subtype 8* appears to be healthy tissue (positively correlated with FEV_1).

Finally, in Section 3.3, we visualize the subtypes on the subject and population levels and explain the clinical interpretation of each subtype. We further justify the discovered subtypes by studying the genetic heritability of each subtype.

3.1. Setup

Feature Extraction Pipeline We apply our method to lung CT inspiratory images of 7,292 subjects from the COPDGene study (Regan et al., 2011). We first segment the lung volume into spatially homogeneous regions that align with image boundaries using the SLIC superpixel segmentation algorithm (Holzer and Donner, 2014). Then for each 3D superpixel, we extract three different types of imaging features that previously have been shown to be important in characterizing emphysema (Shaker et al., 2010; Sorensen et al., 2012): (1) 32-bin intensity histogram features (`Hist`) following Sorensen et al. (2012), (2) Haralick features (`Hara`) that encode image texture but also incorporate intensity (Vogl et al., 2014), and (3) a rotationally invariant descriptor (`sHOG`) proposed by Liu et al. (2014) which computes the histogram of gradients of pixels on a unit sphere using spherical harmonics.

To construct a subject-level representation from the superpixel features, we assume the local features of subject s are samples drawn from a probability distribution p_s. To compute the distribution embedding for each subject as our subject-level representation, we estimate pairwise similarity between subjects' distributions using KL-divergence. However, to avoid imposing any kind of parametric assumptions for KL estimation, we use the nonparametric

KL estimation approach proposed by Schabdach et al. (2017). Our distribution embedding pipeline is described in detail in Appendix A.

Clinical Measurements To evaluate our subject-level representation, we use the representation to predict a few clinical variables that are indicative of disease severity. More specifically, we use the following measurements:

- Percent Predicted Forced Expiratory Volume in one second (FEV$_1$ PP): A measure of lung function which is the percentage of normal predicted values of FEV$_1$ for individuals in the population with similar age, height, weight, gender and ethnicity. Lower values indicate more severe disease.
- Ratio of FEV$_1$ to Forced Vital Capacity (FEV$_1$/FVC): Forced Vital Capacity (FVC) is the total amount of air an individual can exhale forcefully after taking the deepest breath possible. This ratio represents the proportion of an individual's vital capacity that they can breathe out in one second.
- Global Initiative for Obstructive Lung Disease (GOLD): GOLD is a discrete value between zero and four, which is derived from two Spirometry measurements. Zero is used for people at risk (Normal Spirometry but Chronic Symptoms), and 1-4 denote Mild to Very Severe COPD. In this paper, a score of -1 is used for subjects who have Preserved Ratio Impaired Spirometry (PRISm), which indicates that they have reduced FEV$_1$ while having preserved FEV$_1$/FVC.
- Distance Walked: The distance walked in 6 minutes that has been shown to be a good indicator of disease severity in COPD patients (Dajczman et al., 2015).

We report R^2 when evaluating the performance with respect to our continuous measurements (*i.e.*, FEV$_1$ PP, FEV$_1$/FVC, and Distance Walked). For GOLD, which is a discrete but ordered measurement, we report accuracy and also the percentage of cases whose classification lay within one class of the true value (one-off) as well as exact value.

3.2. Quantitative Evaluation of the Subtypes

In this section, we first show that our extracted features are informative by comparing their predictive performance with that of a set of baselines. Next, we show incorporating these features in our variational posterior approximation can improve the performance of generative models. For the details of hyper-parameter setting and additional experiments, including the sensitivity analysis with respect to the number of topics K see Appendix D.

Baselines For each task mentioned above we have a set of baselines. For evaluating the predictive performance of our extracted features, we compare our method with two baselines:

1. Low Attenuation Area below Hounsfield Unit of -950 on Inspiration CT image (%LAA-950Insp) which is commonly used as a clinical measure of emphysema.
2. A subject-level representation learned by a traditional bag-of-words (BOW) model which is the $K-$means algorithm.

We compare the discriminative performances of the three local image descriptors (*i.e.,* `Hara`, `Hist`, `Hist+sHOG`) along with two methods of building the subject-level representation (*i.e.,* $K-$means and our Distribution Distance (KL) method). We separately train linear regression models (via Ridge Regression) to predict FEV_1 PP and FEV_1/FVC from the subject-level features ($\boldsymbol{t}_s$). We use the predicted values to compute the GOLD score[2].

To evaluate the effect of incorporating these features in a generative model via our encoder-decoder framework, we compare our method with two baselines:

1. Topic model with Gaussian observations: Note that the supervised topic models discussed in Section 1 are proposed for documents with discrete observations; hence, we need to devise a topic model baseline that can handle gaussian likelihood and is comparable to our model. We choose Gaussian LDA (G-LDA) model (Das et al. (2015)) as our unsupervised topic model baseline.

2. Supervised topic model with Gaussian observations: We modify G-LDA model (Das et al. (2015)) in a way that it can generate the disease severity y_s given the per-subject subtype proportions π_s. More concretely, we assume $y_s \sim \mathcal{N}(\boldsymbol{\mu}(\pi_s), \sigma^2)$ where $\boldsymbol{\mu}$ is a learnable function and σ^2 is a hyperparameter.

After training the models, we compute the posterior mean of the subtype proportion (*i.e.,* $\mathbb{E}_q[\boldsymbol{\pi}_s|\mathcal{D}]$) on the test data for evaluation. These values are used to train linear regression models predicting the disease severity measures.

Predictive Power of the Representation Table 2 demonstrates our approach outperforms the threshold-based approach (%LAA-950Insp) as well as BOW across all choices of local image descriptors. While all three choices of local image descriptors perform equally well when used by our method, there is significant variation in performances when BOW is used. In the rest of the experiments, we opt to use `Hist+sHOG` as the local image features for computing the subject-level representation due to the slight advantage in performance.

Evaluation of our encoder-decoder framework The results in Table 3 show that our subject-level features, $\boldsymbol{t}_s$, outperform or perform on par with the baselines. The G-LDA, *without* subject-level features $\boldsymbol{t}_s$, learns subtypes that are not predictive of disease severity. Furthermore, the supervised G-LDA, improves the results but still does not perform as well as our approach. Our method and G-LDA baseline converge to ELBO values of 363.95 ± 0.63 and 364.64 ± 0.35 correspondingly. That is, despite outperforming G-LDA in terms of predictive performance, the ELBO in our method is not significantly worse than that of G-LDA. ELBO is computed on the holdout set and is averaged over 5 runs.

3.3. Clinical interpretation

Population-Level Interpretation To summarize the results of the topic model, we compute the posterior distribution of z_{sn}. The $P(z_{sn} = k|\mathcal{D})$ represents the posterior probability of supervoxel n of subject s being assigned to subtype k which can be visualized as a label mask. Examples of such masks are shown in Fig. 2(b) for a few subjects and subtypes.

[2] We pass the predicted values for these two quantities to a learned decision tree classifier to compute GOLD score.

Local Image Feature	Subject-level Descriptor	Exact Acc (Std dev)	One-off Acc (Std dev)
Baseline	%Low Attenuation Level (-950)	0.56 (0.03)	0.76 (0.02)
`Hara`	BOW (K-means)	0.47 (0.02)	0.71 (0.02)
	Distribution Distance (KL)	0.58 (0.03)	0.83 (0.02)
`Hist`	BOW (K-means)	0.54 (0.04)	0.79 (0.01)
	Distribution Distance (KL)	0.57 (0.03)	0.82 (0.01)
`Hist+sHOG`	BOW (K-means)	0.57 (0.03)	0.82 (0.01)
	Distribution Distance (KL)	**0.59** (0.03)	**0.84** (0.01)

Table 2: Average classification accuracy of predicting GOLD 5 classes from subject-level descriptors. Subject-level descriptors are computed from corresponding local image features in each row. `Hara`, `Hist`, `Hist+sHOG` denote Haralick, Histogram, Histogram combined with Spherical Histogram of Gradient descriptors respectively. Results are averaged across 5 cross-validation folds. *One-off Acc* is the percentage of times the predictor was at most one-off in predicting GOLD score. We use Distribution Distance (KL) with `Hist+ sHOG` features as our subject-level descriptor for the rest of experiments.

	R^2			
Subject-Level Descriptor	FEV_1 PP	FEV_1/FVC	FVC	Distance Walked
%Low Attenuation Level (-950)	0.44	0.61	0.03	0.07
BOW (K-means)	0.55	0.66	**0.48**	0.19
G-LDA (Das et al. (2015))	0.35	0.49	0.13	0.12
Supervised G-LDA	0.34	0.51	0.13	**0.21**
Proposed Method ($\boldsymbol{t}_s$)	**0.58**	**0.69**	0.38	0.20
Subject2vec (Singla et al., 2018)	0.68	0.71	-	-

Table 3: Performance of predicting FEV_1 PP, FEV_1/FVC, FVC, and distance walked compared across BOW, G-LDA, supervised G-LDA, our method ($\boldsymbol{t}_s$), and *% Low Attenuation Level (-950) (classic)* subject-level descriptors using ridge regression. Our method outperforms the *G-LDA* and *Supervised G-LDA* in almost all metrics. For *G-LDA*, we use topic proportions inferred by the topic model (Das et al., 2015). *Supervised G-LDA* is a supervised variant of the model proposed by Das et al. (2015) which assumes the disease severity y_s depends on the subtype proportions π_s of subject s. *Subject2Vec* (Singla et al., 2018) is added as a powerful supervised model for reference and an upper bound of performance. The results for FVC and distance walked are not reported by Singla et al. (2018).

We register the label masks of all the subtypes to a common space to compute the average distribution of each subtype across the population. Fig. 2(a) shows these average distributions for each subtype along with corresponding scatter plots denoting the correlation between the proportion of the subtype and FEV_1 PP. Each dot in the scatter plot denotes one subject where $y-$axis corresponds to FEV_1 PP and $x-$axis is the average of the probabilities of that subtype over all supervoxels of the subject. A positive correlation suggests that tissue type is healthy and negative correlation suggests a disease-related subtype.

We also study the average distributions of the subtypes and their variations among patients with different GOLD scores. The result is shown in Fig. 3. Each bar represents a sub-population of patients with a particular GOLD score and colors within the bar repre-

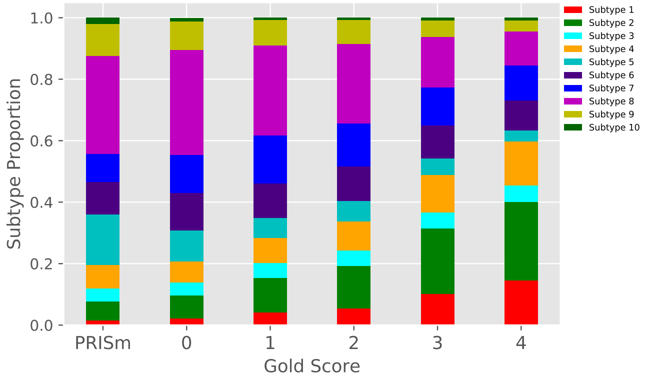


Figure 3: Subtype proportions averaged over subsets of the population with GOLD score values PRISm, 0, 1, 2, 3, and 4.

sent the average proportion of a subtype within that sub-population. All bars have equal sizes but the proportion of subtypes varies. The proportion of *subtype 1* and *2* increase as we move from PRISm to GOLD score 4 (indicating severely diseased). *Subtype 8*, in contrast, decreases with increased severity. *Subtype 5* is notable because even though it is not significantly correlated with disease, it is prevalent in PRISm sub-population relative to other GOLD scores.

Patient-Level Interpretation To have a better understanding of subtypes, we visualize $P(z_{sn} = k|\mathcal{D})$ on lung CT's of nine subjects for $k = 1, 2, 8$ which have the strongest correlation with FEV_1. Fig. 2(b) shows that *subtype 1* is found primarily on pulmonary bullae and *subtype 2* captures patients with peripheral bronchiolitis in patients with severe pulmonary disease (*i.e.,* Gold score ≥ 3). On the other hand *subtype 8* is very pronounced on the rind of three subjects with healthy lungs.

To get a clinical understanding of these subtypes we asked a clinical expert to inspect all subtypes showing average and subject-level representation. Tissue subtypes 1, 2, 3, 4, and 10 are negatively correlated with FEV_1 PP. Thus these subtypes are correlated with increased disease severity. Tissue *subtype 1* tends to characterize paraseptal emphysema and is often found in regions containing pulmonary bullae. *Subtype 1* tends to pick up low attenuation areas on the surface. *Subtype 2* is often indicative of peripheral bronchiolitis, picking up peripheral rind linear opacities in the lung, in some cases blood vessels or lymphatics, as well as tree-in-bud opacities. *Subtype 3* predominantly captures different pathological features. It is associated mostly with large high attenuation areas like scarring and vessels as well as airways. *Subtype 4* picks up on more preserved (*i.e.,* less destruction) areas in patients with emphysema. *Subtype 10* is mostly related to the unexplained image statistics associated with large high attenuation areas.

In contrast subtypes 5, 6, 7, 8, and 9 are negatively correlated with increased disease severity. *Subtype 5* captures regions that are more relatively hyperattenuated than surrounding regions. *Subtype 6* picks up on some dimensional feature of the thorax, maintaining a distance on structure – though it is not clear what it is picking up. This is also true for *subtype 7*, which was difficult for the clinical expert to characterize. Subtypes 5, 6, and 7 tend to be attenuation agnostic. *Subtype 8* is associated with more normal and

Subtype	h^2 (%)	SE (%)	p-value
1	**23.69**	**8.42**	**2.3e-03**
2	**23.37**	**8.29**	**1.8e-03**
3	5.83	7.92	2.2e-01
4	9.96	8.26	1.1e-01
5	≈ 0	8.17	5e-01
6	≈ 0	8.38	5e-01
7	8.37	8.48	1.7e-01
8	**18.74**	**8.34**	**1.1e-02**
9	1.46	8.00	4.3e-01
10	2.16	8.00	3.9e-01

Table 4: Heritability of tissue subtypes. h^2 measures the fraction of phenotypic variance (*i.e.*, variance in subject subtype proportion) explained by the total genetic variance. We denote standard error by SE.

blotchy regions on the rind of the lung. *Subtype 9* is characteristic of thicker peripheral opacities and lines on the apex of the lung which might be indicative of higher diffusing capacity.

Genetic Heritability To understand the genetic etiology of each subtype, we perform the genetic heritability analysis. In brief, the genetic heritability analysis studies the correlation between a quantitive trait and genetic data by estimating the proportion of the variance explained by genetic random effects. The variance ratio (h^2) is estimated under a linear mixed effect model where the fixed effects are nuisance variables, and the random effect is the linear effect of the genotyped variants. The higher the h^2, the stronger the genetic contribution to the trait. For each subtype, we view the proportion as a quantitive trait and estimate h^2 using the Restricted Maximum Likelihood (REML) method using GCTA software (Yang et al., 2010). We use age, gender, number of smoking packs per year, and the first six principal components of the genetic kinship matrix as nuisance parameters (fixed effect). The results are shown in Table 4. *Subtype 1*, *2*, and *8* show significant heritability of approximately $18 - 24\%$, providing strong evidence that these subtypes are biologically driven. While *subtypes 1, 2* have the strongest negative correlation with FEV_1, *subtype 8* has the strongest positive correlation with the FEV_1.

4. Discussion and Conclusion

In this paper, we proposed an approach which lets the practitioner incorporate the predictive features into the posterior approximation of a generative model which is more amenable to interpretation. We showed an application of our method to COPD, which is a highly heterogeneous disease. We viewed every patient as a mixture of different subtypes; hence, a topic model is a proper generative model.

We showed that one could incorporate the discriminative information into the space of the posterior distributions to avoid loss of predictive performance. The idea is that the predictive model shares covariates relevant to prediction ($\boldsymbol{t}_s$) with the generative model. Therefore, they have similar predictive performance. We incorporate $\boldsymbol{t}_s$ into the approximation of the latent variable's posterior distribution. To make the inference computationally efficient, we presented a specific transformation of $\boldsymbol{t}_s$ that results in a closed-form parameterization of the posterior distribution of the subtype proportion.

We apply our model on CT images of the COPDGene dataset. We first demonstrate that our predictive features are more effective for disease severity prediction compared to the standard K-means method. Table 2 shows that our approach achieves the best predictive performance regardless of the input local image descriptor while there is significant variation in the performance of K-means. Furthermore, we show that our framework can outperform unsupervised and supervised topic models. Table 3 shows that the vanilla topic modeling, which is fully unsupervised, completely loses discriminative power. Making the topic model supervised by incorporating the disease severity metrics directly into the generative model, improves the performance but this supervised topic model still underperforms compared to our approach.

The posterior probability of the different latent random variables in our model provides insight into the disease. Figs. $2(a)$ and $2(b)$ visualize the population-level and subject-level distributions of the subtypes. However, not all inferred subtypes are aligned with the current clinical understanding of the disease (*e.g.,* subtypes six, seven, and ten). The fact that subtype ten is positively correlated with FEV_1 suggests that it represents healthy tissue. We observed that the proportion of subtype five is higher in the PRISm sub-population than the rest of the population (Fig. 3). This is a promising area for further investigation since the PRISm patients are difficult to characterize. However, this subtype does not show a significant correlation with the genetic data. Interestingly, the most significant subtypes in term of genetic heritability are the ones with the strongest correlation with FEV_1. Note that in a truncated HDP model there is no guarantee that we find all the subtypes and all subtypes are interpretable (Miller and Harrison, 2014). One of the main motivations for our work was using an informative covariate to ensure the discriminative information is not lost and implicitly having more ”relevant” subtypes. Understanding the biological etiology of these subtypes requires further causal analysis, which is another avenue for future research.

References

ZAA. Aziz, Athol U Wells, David M. Hansell, Gordon A. Bain, Susan Jennifer Copley, Sujal R. Desai, Stephen M. Ellis, Fergus Vincent Gleeson, Suzana Grubnic, Andrew G. Nicholson, Simon P. G. Padley, Kate S Pointon, John Hughes Reynolds, Rowena Robertson, and MichaelB. Rubens. Hrct diagnosis of diffuse parenchymal lung disease: inter-observer variation. *Thorax*, 59 6:506–11, 2004.

Nematollah K Batmanghelich, Ardavan Saeedi, Michael Cho, Raul San Jose Estepar, and Polina Golland. Generative method to discover genetically driven image biomarkers. In *International Conference on Information Processing in Medical Imaging*, pages 30–42. Springer, 2015.

Polina Binder, Nematollah K. Batmanghelich, Raul San Jose Estepar, and Polina Golland. Unsupervised Discovery of Emphysema Subtypes in a Large Clinical Cohort. In Li Wang, Ehsan Adeli, Qian Wang, Yinghuan Shi, and Heung-Il Suk, editors, *Machine Learning in Medical Imaging: 7th International Workshop, MLMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings*, pages 180–187. Springer International Publishing, Cham, 2016. ISBN 978-3-319-47157-0. doi: 10.1007/978-3-319-47157-0{_}22. URL http://link.springer.com/10.1007/978-3-319-47157-0_22.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 1533-7928.

Michael Bryant and Erik B. Sudderth. Truly nonparametric online variational inference for hierarchical dirichlet processes. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, pages 2699–2707, USA, 2012. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999325.2999436.

Peter J Castaldi, Marta Benet, Hans Petersen, Nicholas Rafaels, James Finigan, Matteo Paoletti, H Marike Boezen, Judith M Vonk, Russell Bowler, Massimo Pistolesi, et al. Do copd subtypes really exist? copd heterogeneity and clustering in 10 independent cohorts. *Thorax*, 72(11):998–1006, 2017a.

Peter J Castaldi, Marta Benet, Hans Petersen, Nicholas Rafaels, James Finigan, Matteo Paoletti, H Marike Boezen, Judith M Vonk, Russell Bowler, Massimo Pistolesi, Milo A Puhan, Josep Anto, Els Wauters, Diether Lambrechts, Wim Janssens, Francesca Bigazzi, Gianna Camiciottoli, Michael H Cho, Craig P Hersh, Kathleen Barnes, Stephen Rennard, Meher Preethi Boorgula, Jennifer Dy, Nadia N Hansel, James D Crapo, Yohannes Tesfaigzi, Alvar Agusti, Edwin K Silverman, and Judith Garcia-Aymerich. Do copd subtypes really exist? copd heterogeneity and clustering in 10 independent cohorts. *Thorax*, 72(11):998–1006, 2017b. ISSN 0040-6376. doi: 10.1136/thoraxjnl-2016-209846. URL https://thorax.bmj.com/content/72/11/998.

George H Chen and Jeremy C Weiss. Survival-supervised topic modeling with anchor words: Characterizing pancreatitis outcomes. *arXiv preprint arXiv:1712.00535*, 2017.

Irene Y Chen, Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. Probabilistic machine learning for healthcare. *arXiv preprint arXiv:2009.11087*, 2020.

Xu Chen, Xiaomao Xu, and Fei Xiao. Heterogeneity of chronic obstructive pulmonary disease: from phenotype to genotype. *Frontiers of medicine*, 7(4):425–32, 12 2013. doi: 10.1007/s11684-013-0295-x. URL http://www.ncbi.nlm.nih.gov/pubmed/24234678.

Angel Cruz-Roa, Gloria Díaz, Eduardo Romero, and Fabio A González. Automatic annotation of histopathological images using a latent topic model based on non-negative matrix factorization. *Journal of pathology informatics*, 2, 2011.

Esther Dajczman, Rima Wardini, Goulnar Kasymjanova, David Préfontaine, Marc Alexander Baltzan, and Norman Wolkove. Six minute walk distance is a predictor of survival in patients with chronic obstructive pulmonary disease undergoing pulmonary rehabilitation. *Canadian respiratory journal*, 22(4):225–229, 2015.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804, 2015.

Marc Decramer, Wim Janssens, and Marc Miravitlles. Chronic obstructive pulmonary disease. *The Lancet*, 379(9823):1341 – 1351, 2012. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(11)60968-9. URL http://www.sciencedirect.com/science/article/pii/S0140673611609689.

A. Depeursinge, D. Sage, A. Hidki, A. Platon, P. Poletti, M. Unser, and H. Muller. Lung tissue classification using wavelet frames. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6259–6262, Aug 2007. doi: 10.1109/IEMBS.2007.4353786.

Rick Farouni. A contemporary overview of probabilistic latent variable models. *arXiv preprint arXiv:1706.08137*, 2017.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531 vol. 2, June 2005. doi: 10.1109/CVPR.2005.16.

Lauren A Hannah, David M Blei, and Warren B Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 12(6), 2011.

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.

M Holzer and R Donner. Over-Segmentation of 3D Medical Image Volumes based on Monogenic Cues. *Cvww*, (JANUARY 2014):35–42, 2014.

Michael C Hughes, Gabriel Hope, Leah Weiner, Thomas H McCoy Jr, Roy H Perlis, Erik B Sudderth, and Finale Doshi-Velez. Semi-supervised prediction-constrained topic models. In *AISTATS*, pages 1067–1076, 2018.

Y. Häme, E. D. Angelini, M. A. Parikh, B. M. Smith, E. A. Hoffman, R. G. Barr, and A. F. Laine. Sparse sampling and unsupervised learning of lung texture patterns in pulmonary emphysema: Mesa copd study. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 109–113, April 2015. doi: 10.1109/ISBI.2015.7163828.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, Nov 1999. ISSN 1573-0565. doi: 10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint*, (Ml):1–14, 2013. URL http://arxiv.org/abs/1312.6114.

Iryna Korshunova, Hanchen Xiong, Mateusz Fedoryszak, and Lucas Theis. Discriminative topic modeling with logistic lda. In *Advances in Neural Information Processing Systems*, pages 6767–6777, 2019.

Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904, 2009.

Yue Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Yan Miao, Weiqi Liu, Tamas Ordog, Joanna M Biernacka, et al. Inferring multimodal latent topics from electronic health records. *Nature communications*, 11(1):1–17, 2020.

Scott W. Linderman, Matthew J. Johnson, and Ryan P. Adams. Dependent multinomial models made easy: Stick breaking with the pólya-gamma augmentation. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3456–3464, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969442.2969625.

Kun Liu, Henrik Skibbe, Thorsten Schmidt, Thomas Blein, Klaus Palme, Thomas Brox, and Olaf Ronneberger. Rotation-Invariant HOG Descriptors Using Fourier Analysis in Polar and Spherical Coordinates. *International Journal of Computer Vision*, 106(3):342–364, 2014. ISSN 09205691. doi: 10.1007/s11263-013-0634-z.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112, 01 2016. doi: 10.1080/01621459.2017.1285773.

Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

Jeffrey W Miller and Matthew T Harrison. Inconsistency of pitman-yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370, 2014.

Yang Shin Park, Joon Beom Seo, Namkug Kim, Eun Jin Chae, Yeon Mok Oh, Sang Do Lee, Youngjoo Lee, and Suk-Ho Kang. Texture-based quantification of pulmonary emphysema on high-resolution computed tomography: comparison with density-based quantification and correlation with pulmonary function test. *Investigative radiology*, 43 6:395–402, 2008.

C. Peterson and J. R. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

Emma Pierson, Pang Wei Koh, Tatsunori Hashimoto, Daphne Koller, Jure Leskovec, Nick Eriksson, and Percy Liang. Inferring multidimensional rates of aging from cross-sectional data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 97–107. PMLR, 2019.

Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad. Learning probabilistic phenotypes from heterogeneous ehr data. *Journal of biomedical informatics*, 58: 156–165, 2015.

Mithun Prasad, Arcot Sowmya, and Peter Wilson. Multi-level classification of emphysema in hrct lung images. *Pattern Analysis and Applications*, 12(1):9–20, Feb 2009. ISSN 1433-755X. doi: 10.1007/s10044-007-0093-7. URL https://doi.org/10.1007/s10044-007-0093-7.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2011.

Jason Ren, Russell Kunes, and Finale Doshi-Velez. Prediction focused topic models via vocab selection. *arXiv preprint arXiv:1910.05495*, 2019.

James Clark Ross, Peter J. Castaldi, Michael H. Cho, Junxiang Chen, Yale Chang, Jennifer G. Dy, Edwin K. Silverman, George R. Washko, and Raúl San José Estépar. A bayesian nonparametric model for disease subtyping: Application to emphysema phenotypes. *IEEE Transactions on Medical Imaging*, 36:343–354, 2016.

Jenna Schabdach, William M Wells, Michael Cho, and Kayhan N Batmanghelich. A likelihood-free approach for characterizing heterogeneous diseases in large-scale studies. In *International Conference on Information Processing in Medical Imaging*, pages 170–183. Springer, 2017.

Saher B Shaker, Marleen De Bruijne, Lauge Sorensen, Saher B Shaker, and Marleen De Bruijne. Quantitative analysis of pulmonary emphysema using local binary patterns. *Medical Imaging, IEEE Transactions on*, 29(2):559–569, 2010.

S D Shapiro. Evolving concepts in the pathogenesis of chronic obstructive pulmonary disease. *Clin Chest Med*, 21(4):621–632, 2000.

Sumedha Singla, Mingming Gong, Siamak Ravanbakhsh, Frank Sciurba, Barnabas Poczos, and Kayhan N Batmanghelich. Subject2vec: generative-discriminative approach from a set of image patches to a vector. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 502–510. Springer, 2018.

Jingkuan Song, Jie Yang, Benjamin M. Smith, Pallavi P. Balte, Eric A. Hoffman, Richard G Barr, Andrew F. Laine, and Elsa D. Angelini. Generative method to discover emphysema subtypes with unsupervised learning using lung macroscopic patterns (lmps): The mesa copd study. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 375–378, 2017.

L. Sorensen, S. B. Shaker, and M. de Bruijne. Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE Transactions on Medical Imaging*, 29(2):559–569, Feb 2010. ISSN 0278-0062. doi: 10.1109/TMI.2009.2038575.

Lauge Sorensen, Mads Nielsen, Pechin Lo, Haseem Ashraf, Jesper H. Pedersen, and Marleen De Bruijne. Texture-based analysis of COPD: A data-driven approach. *IEEE Transactions on Medical Imaging*, 31 (1):70–78, 2012. ISSN 02780062. doi: 10.1109/TMI.2011.2164931.

Mani Sotoodeh and Joyce C Ho. Improving length of stay prediction using a hidden markov model. *AMIA Summits on Translational Science Proceedings*, 2019:425, 2019.

Y W Teh, M I Jordan, M J Beal, and D M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Renuka Uppaluri, Theophano Mitsa, Milan Sonka, EricA. Hoffman, and Geoffrey McLennan. Quantification of pulmonary emphysema from lung computed tomography images. *American Journal of Respiratory and Critical Care Medicine*, 156(1):248–254, 1997. doi: 10.1164/ajrccm.156.1.9606093. PMID: 9230756.

Iñigo Urteaga, Mollie McKillop, and Noémie Elhadad. Learning endometriosis phenotypes from patient-generated data. *NPJ digital medicine*, 3(1):1–14, 2020.

Filippo Valle, Matteo Osella, and Michele Caselle. A topic modeling analysis of tcga breast and lung cancer transcriptomic data. *Cancers*, 12(12):3799, 2020.

G. Viegi, F. Pistelli, D. L. Sherrill, S. Maio, S. Baldacci, and L. Carrozzi. Definition, epidemiology and natural history of copd. *European Respiratory Journal*, 30(5):993–1013, 2007. ISSN 0903-1936. doi: 10.1183/09031936.00082507. URL https://erj.ersjournals.com/content/30/5/993.

Wolf Dieter Vogl, Helmut Prosch, Christina Muller-Mang, Ursula Schmidt-Erfurth, and Georg Langs. Longitudinal alignment of disease progression in fibrosing interstitial lung disease. In *Lecture Notes in Computer Science*, volume 8674 LNCS, pages 97–104, 2014. ISBN 9783319104690. doi: 10.1007/978-3-319-10470-6{_}13.

Simon LF Walsh, Lucio Calandriello, Mario Silva, and Nicola Sverzellati. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *The Lancet Respiratory Medicine*, 6(11):837–845, 2018.

World Health Organization. The top 10 causes of death. https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death, 5 2018. [Online; accessed 12-June-2018].

Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, and Others. Common {SNPs} explain a large proportion of the heritability for human height. *Nat Gen*, 42(7):565–569, 2010. ISSN 1546-1718. doi: 10.1038/ng.608.Common.

Jie Yang, Elsa D. Angelini, Pallavi P. Balte, Eric A. Hoffman, John H. M. Austin, Benjamin M. Smith, Jingkuan Song, Richard G Barr, and Andrew F. Laine. Unsupervised discovery of spatially-informed lung texture patterns for pulmonary emphysema: The mesa copd study. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 10433:116–124, 2017.

Jie Yang, Xinyang Feng, Andrew F Laine, and Elsa D Angelini. Characterizing alzheimer's disease with image and genetic biomarkers using supervised topic models. *IEEE Journal of Biomedical and Health Informatics*, 24(4):1180–1187, 2019.

Incorporating External Information in Tissue Subtyping: A Topic Modeling Approach (Supplementary Material)

Ardavan Saeedi[*][†] AV.SAEEDI@GMAIL.COM
Hyperfine

Payman Yadollahpour[*] PYADOLLA@BROADINSTITUTE.ORG
Broad Institute

Sumedha Singla SUMEDHA.SINGLA@PITT.EDU
University of Pittsburgh

Brian Pollack BRP98@PITT.EDU
University of Pittsburgh

William Wells SW@BWH.HARVARD.EDU
Harvard Medical School / Brigham and Women's Hospital

Frank Sciurba CIURBAFC@UPMC.EDU
University of Pittsburgh Medical Center

Kayhan Batmanghelich KAYHAN@PITT.EDU
University of Pittsburgh

Appendix A: Pipeline for Constructing the Subject-level Features

In this section, we provide an overview of our feature-extraction pipeline for a disease severity prediction task. Consider a discriminative model for predicting disease severity y_s from a subject's lung CT image I_s. We define this model as a composition of two functions: (1) $f(\cdot)$ which is a function that extracts local descriptors from image I_s, hence $\mathcal{X}_s = f(I_s)$, and (2) an aggregation function, $g(\cdot)$ which we use to construct subject-level features relating the subject to the rest of the population. We minimize

$$\ell(y_s; h(\overbrace{g(f(I_s))}^{t_s})), \tag{1}$$

where h is a regressor or a classifier, depending on y being continuous or discrete and $\ell(\cdot;\cdot)$ is a loss function that is chosen accordingly. We define $\mathbf{t}_s \triangleq g(\mathcal{X}_s)$ to be the features relating the subject to the rest of the population. Each of the functions can either be hand engineered or learned; for example $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ can consist of different layers of a CNN, or a combination of hand engineered feature functions with aggregation performed

[*] Equal contribution
[†] Work is not related to the research done at Hyperfine.

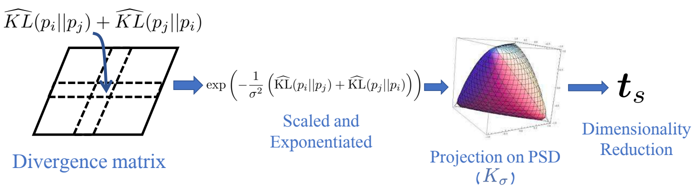

$\widehat{KL}(p_i||p_j)+\widehat{KL}(p_j||p_i)$
Divergence matrix
$\exp\left(-\frac{1}{\sigma^2}\left(\widehat{\mathrm{KL}}(p_i\|p_j)+\widehat{\mathrm{KL}}(p_j\|p_i)\right)\right)$
Scaled and Exponentiated
Projection on PSD (K_σ)
$\boldsymbol{t}_s$
Dimensionality Reduction

Figure 1: Construction of the subject-level features ($\boldsymbol{t}_s$) has the following steps: approximating pairwise divergence matrix, exponentiating the matrix, projecting it on the PSD cone, and reducing the dimensionality.

by summation, followed by prediction via a regression model. In this paper, $f(\cdot)$, is a hand-crafted feature extractor but the same machinery applies to deep learning based features.

We model local features of subject s as samples drawn from its probability distribution p_s. The aggregator maps the probability density to a vector $\boldsymbol{t}_s$ relating the subject to the rest of the population. To do that, we take the following steps. First, we estimate the Kullback-Leibler (KL) divergence between every pair of probability distributions. Second, we convert the distribution distance to a proper similarity kernel. Finally, we use a dimensionality reduction method to estimate $\boldsymbol{t}_s$ from the similarity kernel. The pipeline is shown in Fig. 1.

Estimating KL divergence The KL divergence has the following form,

$$\mathrm{KL}(p_i\|p_j)=\int_{\mathbb{R}^d}\log\frac{p_i(x)}{p_j(x)}p_i(x)dx. \tag{2}$$

In this section, we do not assume any explicit parametric form for p_i. Even with a parametric form, estimating the KL divergence is not straightforward. Instead of assuming an explicit parametrization, we use a non-parametric estimator for KL divergence that is consistent and unbiased (Poczos and Schneider, 2011). The estimator is scalable for high-dimensional features and it only requires the nearest neighbor graph that can be approximated using a hashing method (Schabdach et al., 2017). We use $\widehat{\mathrm{KL}}(p_i\|p_j)$ to denote the estimator for the KL divergence.

Computing the Similarity Kernel Matrix The similarity kernel matrix is a Positive Semi-Definite (PSD) matrix. For example, exponentiating the ℓ_2-distance between features results in a proper similarity kernel matrix known as an RBF kernel. However, the KL divergence is neither symmetric nor a proper metric. First, we compute an $S\times S$ matrix where the entry in row i and column j is

$$[L_\sigma]_{ij}=\exp\left(-\frac{1}{\sigma^2}\left(\widehat{\mathrm{KL}}(p_i\|p_j)+\widehat{\mathrm{KL}}(p_j\|p_i)\right)\right). \tag{3}$$

The variable σ is set to the median of KL divergences (so-called median trick (Song et al., 2010)). Then, we project this matrix onto the PSD cone to construct the kernel,

$$K_\sigma=\mathrm{Proj}_{\mathrm{PSD}}(L_\sigma), \tag{4}$$

where $\mathrm{Proj}_{\mathrm{PSD}}$ computes the Singular Value Decomposition of the input matrix and sets the negative singular values to zero.

Computing Subject Representation ($\boldsymbol{t}_s$) Since K_σ is a PSD matrix, one can compute $K_\sigma = BB^T$ and view columns of B as an *implicit* characterization of the subjects. However, the columns of B are high dimensional (as many as the number of patients in the dataset). We use Locally Linear Embedding (LLE) to reduce the dimensionality (Zhang and Wang, 2006). Other dimensionality reduction methods can be applied as well.

Appendix B: Non-parametric inference of the divergence

The Kullback-Leibler (KL) divergence between probability densities p_i and q_j is defined as follows:

$$\mathrm{KL}(p_i \| p_j) = \int_{\mathbb{R}^d} \log \frac{p_i(x)}{p_j(x)} p_i(x) dx.$$

Poczos et al. (2011) proposed to estimate the divergences without assuming a parametric form for the probability densities. To avoid making a global parametric assumption for p_i and p_j, they parameterize them locally and use the *local* log-likelihood method (Loader, 1996) to estimate the local parameters.

Let us assume that $S_i = \{x_{i1}, \cdots, x_{iN}\}$ and $S_j = \{x_{j1}, \cdots, x_{jM}\}$ are collections of samples drawn from p_i and p_j respectively. With mild assumptions on the probability density, p_i can be represented as $p_i(x) = \tilde{p}_i(x)/Z_{\tilde{p}_i}$, where $\tilde{p}_i(x)$ is an unknown positive function and $Z_{\tilde{p}_i}$ is the corresponding normalizer (*i.e.,* $Z_{\tilde{p}_i} = \int \tilde{p}_i(x)dx$; if $\tilde{p}_i(x)$ is a probability density, $Z_{\tilde{p}_i} = 1$). The $\log \tilde{p}_i(x)$ can be approximated using a polynomial expansion around x, namely

$$\log \tilde{p}_i(u)|_x \approx a_0 + (u-x)^T a_1 + (u-x)^T a_2 (u-x), \tag{5}$$

where a_0, a_1, a_2 are scalar, vector and matrix parameters, respectively, and vary depending on x. The local log-likelihood of the function $\tilde{p}_i$ at point x is:

$$\mathcal{L}_x(\tilde{p}_i) = \sum_{v \in S_i} w\left(\frac{x-v}{h}\right) \log \tilde{p}_i(v) - |S_i| \int w\left(\frac{y-x}{h}\right) \tilde{p}_i(y) dy, \tag{6}$$

where $w(x) = \mathbb{I}(\|x\| \leq 1)$ is a window function and h is a bandwidth. Since the approximation of $\log \tilde{p}_i(x)$ is locally valid, it is reasonable to keep h small; if h goes to infinity, Eq. (6) amounts to the ordinary likelihood estimation and the last term converges to $|S_i| Z_{\tilde{p}_i}$. Poczos et al. (2011) and others (*e.g.,* Gao et al. (2016)) proposed to use local and adaptive bandwidth, *i.e.,* h is a function of x. A popular choice for h is to set it to the 1−NN distance from x; $h(x) \equiv \rho_{k,S_i}(x) \triangleq \min_{v \in S_i} \|v - x\|_2$ similar to Poczos et al. (2011). Optimizing Eq. (6), we get the following form for $\tilde{p}_i$ Loader (1996),

$$\frac{d\mathcal{L}_x(\tilde{p}_i)}{da_0} = \sum_{v \in S_i} w\left(\frac{x-\psi(v)}{h}\right) - |S_i| \int w\left(\frac{y-x}{h}\right) e^{a_0} dy = 0, \tag{7}$$

$$\tilde{p}_i(x) = \frac{1}{|S_i| h \int w(x) dx} \sum_{v \in S_i} w(v) = \frac{k}{|S_i| C_d \rho^d_{k,S_i}(x)}, \; C_d \equiv \frac{\pi^{d/2}}{\Gamma(d/2+1)}. \tag{8}$$

The $C_d \rho_{k,S_i}^d(x)$ are the volumes of d-dimensional balls with radius of one and $\Gamma(\cdot)$ is the Gamma function. Using the re-substitution, we estimate the KL divergences as follows:

$$\widehat{\mathrm{KL}(p_i \| p_j)} = \frac{d}{|S_i|} \sum_{v \in S_i} \log \frac{\rho_{k,S_i}(v)}{\rho_{k,S_j}(v)} + \log \frac{|S_j|}{|S_i| - 1},$$

The estimators are unbiased and consistent (Poczos et al., 2011). In other words, as the number of patches in S_i and S_j increases, the estimations converge to the true value.

Appendix C: Variational Inference

Update Equations for Variational Parameters of Explainer Model

Recall that the joint likelihood is of the form:

$$p(\mathcal{D}, \mathcal{S}, \mathcal{C}) = p(\boldsymbol{\beta}|\alpha) \prod_k p(\boldsymbol{\theta}_k|\eta) \prod_s p(\boldsymbol{\pi}_s|\boldsymbol{\beta}) \prod_{s,n} p(\boldsymbol{x}_{sn}|z_{sn}, \{\boldsymbol{\theta}_k\}) p(z_{sn}|\boldsymbol{\pi}_s), \tag{9}$$

and we assume the following form for the approximate posterior, $q(\cdot)$,

$$\mathcal{Q}' : q(\mathcal{S}, \mathcal{C}) = q(\boldsymbol{\beta}; \beta^*) \underbrace{\prod_s q(\boldsymbol{\pi}_s|\boldsymbol{t}_s; \boldsymbol{W})}_{\text{subject-level}} \underbrace{\prod_{s,n} q(z_{sn}; \boldsymbol{\varphi}_{sn})}_{\text{local descriptor}} \underbrace{\prod_k q(\boldsymbol{\theta}_k; \boldsymbol{\lambda}_k)}_{\text{population-level}}, \tag{10}$$

where β^*, $\boldsymbol{\varphi}_{sn}$, $\boldsymbol{W}$, and $\boldsymbol{\lambda}_k$ are the variational parameters corresponding to the random variables β, z_{sn}, $\boldsymbol{\pi}_s$, and $\boldsymbol{\theta}_k$ respectively. We used empirical Bayes for β meaning that $q(\boldsymbol{\beta}; \beta^*)$ is modeled as a delta function. The $q(\boldsymbol{\pi}_s|\boldsymbol{t}_s; \boldsymbol{W})$ was explained in the main text. The $\boldsymbol{\theta}_k$ and $\boldsymbol{\beta}$ are the population-level random variables. As mentioned, we assume a conjugate prior for $\boldsymbol{\theta}_k$; hence, the optimal variational distribution $q(\boldsymbol{\theta_k}; \boldsymbol{\lambda_k})$ is also in the same family.

In the following, we provide the update equations for each of the variational parameters.

Update of $\boldsymbol{\lambda}_k$ in $q(\boldsymbol{\theta_k}; \boldsymbol{\lambda}_k)$ We model the θ_k with Normal-inverse-Wishart (NIW) distribution which is an exponential family of distributions. The probability densities of the exponential families can be written as follows,

$$p(\boldsymbol{\theta}; \boldsymbol{\lambda}) = h_{\mathrm{NIW}}(\boldsymbol{\theta}) \exp\left(\boldsymbol{\lambda}^T t_{\mathrm{NIW}}(\boldsymbol{\theta}) - A_{\mathrm{NIW}}(\boldsymbol{\lambda})\right),$$

where $t_{\mathrm{NIW}}(\boldsymbol{\theta})$ is called sufficient statistics of the NIW distribution,

$$A_{\mathrm{NIW}}(\boldsymbol{\lambda}) = \log \int \exp\left(\boldsymbol{\lambda}^T t_{\mathrm{NIW}}(\boldsymbol{\theta})\right) dh(\boldsymbol{\theta})$$

is the log partition function and $h_{\mathrm{NIW}}(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ is called the base measure of the NIW distribution. All $A_{\mathrm{NIW}}(\cdot)$, $t_{\mathrm{NIW}}(\cdot)$, and $h_{\mathrm{NIW}}(\cdot)$ are known functions.

$$\begin{aligned} \boldsymbol{\lambda_k} &\leftarrow (1-\rho)\boldsymbol{\lambda_k} + \rho(\eta_k + m \cdot \tilde{t}_x^k), \\ \tilde{t}_x^k &\triangleq \sum_s \sum_n \mathbb{E}_{q(z_{sn})} \left[\nVdash[z_{sn} = k]\right] t_{\mathcal{N}}^k(x_{sn}), \end{aligned} \tag{11}$$

where ρ is the stepsize, m is minibatch scaling, and $\tilde{t}^k_{\mathcal{N}}$ is the expected sufficient statistics of the Gaussian distribution. $\nVdash[\cdot]$ is 1 if its argument is true and 0 otherwise; $\mathbb{E}_{q(z_{sn})}\left[\nVdash[z_{sn} = k]\right]$ is the posterior expectation of z_{sn} being k. For the detailed derivation of the update equation for the Normal-Inverse-Wishart distribution see for instance Guan et al. (2010).

Update of $\boldsymbol{\beta}^*$ in $q(\boldsymbol{\beta}; \boldsymbol{\beta}^*)$ We model the $p(\boldsymbol{\beta}; \gamma)$ as a beta distribution. Similar to Johnson and Willsky (2014), we use a point estimate for $q(\beta)$: $q(\beta) = \delta_{\beta^*}(\beta)$. We use gradient descent to find $\boldsymbol{\beta}^*$. The gradient of $\boldsymbol{\beta}$ is computed as follows,

$$\nabla_{\beta^*}\mathcal{L} = \nabla_{\beta^*}\left\{\mathbb{E}_{q(\boldsymbol{\pi}_s|\boldsymbol{t}_s;\boldsymbol{W})}\left[\ln\frac{p(\beta,\pi)}{q(\beta)q(\pi)}\right]\right\} = \nabla_{\beta^*}\left\{\ln p(\beta^*;\gamma) + \sum_s \mathbb{E}_{q(\boldsymbol{\pi}_s|\boldsymbol{t}_s;\boldsymbol{W})}\left[\ln p(\boldsymbol{\pi}_s|\beta^*)\right]\right\}.$$

We use Monte Carlo sampling to estimate $\mathbb{E}_{q(\boldsymbol{\pi}_s|\boldsymbol{t}_s;\boldsymbol{W})}\left[\ln p(\boldsymbol{\pi}_s|\beta^*)\right]$ by generating samples from $q(\boldsymbol{\pi}_s|\boldsymbol{t}_s;\boldsymbol{W})$. Note that β^* needs to satisfy $\beta^* \geq 0$ after each update.

Update of $\boldsymbol{\varphi}_{sn}$ in $q(z_{sn}; \boldsymbol{\varphi}_{sn})$ Here we derive the variational parameters $\boldsymbol{\varphi}_{sn}$ corresponding to z_{sn}, which is the variable holding the topic assignment to supervoxel n of lung CT of subject s. We follow the standard mean-field approach for the update of this parameter,

$$[\boldsymbol{\varphi}_{sn}]_k \propto \exp\Big[\mathbb{E}_q[\log[\boldsymbol{\pi}_s]_k] + \mathbb{E}_q[\log\mathcal{N}(\boldsymbol{x}_{sn};\boldsymbol{\theta}_k)]\Big],$$

where $[\cdot]_k$ indexes the k^{th} element of the vector. The second term, $\mathbb{E}_q[\log\mathcal{N}(\boldsymbol{x}_{sn};\boldsymbol{\theta}_k)]$, is a standard term that can be find in textbooks about variational inference (Bishop, 2006). However, we need to derive the quantity $\mathbb{E}_q[\log[\boldsymbol{\pi}_s]_k]$, which can be expanded using the stick-breaking construction (Eq. 5 in the main text),

$$\mathbb{E}_q[\log[\boldsymbol{\pi}_s]_k] = \mathbb{E}_q\Big[\log\sigma([\boldsymbol{\psi}_s]_k) + \sum_{j<k}\log\sigma(-[\boldsymbol{\psi}_s]_j)\Big], \tag{12}$$

We can expand the first expectation in Eq. 12,

$$\begin{aligned}
\mathbb{E}_q[\log\sigma([\boldsymbol{\psi}_s]_k)] = \mathbb{E}_{\boldsymbol{\pi}_s}[\log\sigma([\boldsymbol{\psi}_s]_k)] &= \mathbb{E}_{\boldsymbol{\psi}_s\sim\mathcal{N}(\cdot;\boldsymbol{\mu}(\boldsymbol{t}_s;\boldsymbol{W}_{\boldsymbol{\mu}}),diag(\boldsymbol{\sigma}(\boldsymbol{t}_s;\boldsymbol{W}_{\boldsymbol{\sigma}})^2)}\left[\log\sigma([\boldsymbol{\psi}_s]_k)\left|\frac{\partial[\boldsymbol{\pi}_s]_i}{\partial[\boldsymbol{\psi}_s]_j}\right|^{-1}\right] \\
&= \mathbb{E}_{[\boldsymbol{\psi}_s]_k}\left[\log\sigma([\boldsymbol{\psi}_s]_k)\left|\frac{\partial[\boldsymbol{\pi}_s]_i}{\partial[\boldsymbol{\psi}_s]_j}\right|^{-1}\right] \\
&= \mathbb{E}_{[\boldsymbol{\psi}_s]_k}\left[\log\sigma([\boldsymbol{\psi}_s]_k)\left(\prod_{k=1}^{K}\sigma([\boldsymbol{\psi}_s]_k)\prod_{j<k}\sigma(-[\boldsymbol{\psi}_s]_j)\right)^{-1}\right],
\end{aligned} \tag{13}$$

where the expectation can be computed using a Monte Carlo method by sampling from $[\psi_s]_k$. The second term in Eq. 12 can be written analogously.

The update to $\boldsymbol{W}$, the parameters of the subject specific topic proportions, cannot be written in closed form but the gradient of $\mathcal{L}(\cdot)$ with respect to them is,

$$\nabla_{\boldsymbol{W}_i}\mathcal{L}(q) = \sum_{s=1,k=1}^{S,K+1}\nabla_{\boldsymbol{W}_i}\mathbb{E}_q[\log[\boldsymbol{\pi}_s]_k]\Big(([\alpha\boldsymbol{\beta}]_k - 1) + \sum_n^{N_s}[\boldsymbol{\varphi}_{sn}]_k\Big) - \sum_{s=1}^{S}\nabla_{\boldsymbol{W}_i}\mathbb{E}_q[\log q(\boldsymbol{\pi}_s|\boldsymbol{t}_s;\boldsymbol{W})], \tag{14}$$

where we have assumed $\boldsymbol{W}$ is rearranged in vector form for simplified indexing. The expectation in the first term is derived above.

Appendix D: Hyperparameters and additional experiments

Hyperparameters In this section, we provide the initialization setup and the set of hyperparameters used in our experiments. To initialize the parameters of the NIW distribution, $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$, for inference in the topic model, we run unsupervised hierarchical clustering (Campello et al., 2013) on local image features extracted from supervoxels of the training set. The hierarchical clustering cut-off threshold was set to match the number of tissue subtypes K. Each subtype distribution was subsequently initialized with the sufficient statistics computed from the corresponding cluster.

Hyperparameter	Values
α	{1.0, 10.0, 100.0}
γ	{2.0, 10.0, 100.0}
K	{2, 3, 5, 10, 15, 20, 30, 40, 50}
SGD minibatch size	{16, 128}
l^2 regularization	$\{0, 10^{-5}\}$

Table 1: Hyperparameters used in our experiments.

Predictive power of features learned via our encoder-decoder framework Since our model uses $\boldsymbol{t}_s$ for inference, our prediction performance is the same. Our inference algorithm transforms $\boldsymbol{t}_s$ to compute $\mathbb{E}_q[\boldsymbol{\pi}_s|\mathcal{D}]$. If this transformed value is used for the prediction, R^2 of predicting FEV_1 PP and FEV_1/FVC are 0.42 and 0.58 respectively. The gap between these values and the performance of $\boldsymbol{t}_s$ is the cost we pay to gain interpretation, which is much better than the fully unsupervised method. This confirms that our model learns tissue subtypes that are relevant to disease prediction, and not simply capturing irrelevant image statistics in the subject CT's.

Sensitivity to K We investigate the sensitivity of our method to the choice of the number of subtypes, K, which is the most important one amongst the hyperparameters. Fig. 2 shows the results of running the inference for the topic model for varying values of parameter K. We measure the model's ability to explain the observed data (*i.e.,* image features of the lung) on the test set by computing the log-likelihood of the data under the model. Each point is an average over two separate inference runs of the topic model with random initialization. When the assumed number of subtypes is less than 10 the model's performance suffers but for values ≥ 10 we see relatively stable performance. This suggests that our choice of 10 subtypes is a reasonable approximation of the number of image feature clusters.

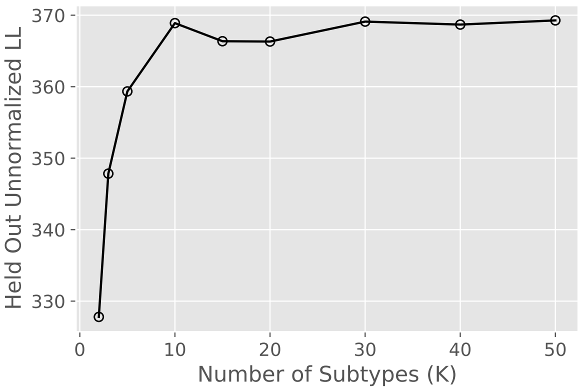


Figure 2: Log-likelihood (LL) of the topic model (with discriminative feature injection) on the held out set for different values of K. Each point is an average over two separate training runs of the model with random initialization.

References

C M Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006. ISBN 0387310738 9780387310732.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

Weihao Gao, Sewoong Oh, and Pramod Viswanath. Breaking the Bandwidth Barrier: Geometrical Adaptive Entropy Estimation. 9 2016. URL http://arxiv.org/abs/1609.02208.

Yue Guan, Jennifer G Dy, Donglin Niu, and Zoubin Ghahramani. Variational inference for nonparametric multiple clustering. In *KDD10 Workshop on Discovering, Summarizing, and Using Multiple Clusterings*, 2010.

Matthew J. Johnson and Alan S. Willsky. Stochastic variational inference for Bayesian time series models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1854–1862, 2014.

Clive R. Loader. Local likelihood density estimation. *Annals of Statistics*, 24(4):1602–1618, 1996. ISSN 00905364. doi: 10.1155/2010/754851.

Barnabas Poczos and Jeff Schneider. On the estimation of alpha-divergences. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 609–617, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL http://proceedings.mlr.press/v15/poczos11a.html.

Barnabás Poczos, Liang Xiong, and Jeff Schneider. Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions. *Uncertainty in Artificial Intelligence*, 2011.

Jenna Schabdach, William M Wells, Michael Cho, and Kayhan N Batmanghelich. A likelihood-free approach for characterizing heterogeneous diseases in large-scale studies. In *International Conference on Information Processing in Medical Imaging*, pages 170–183. Springer, 2017.

Le Song, Sajid M Siddiqi, Geoffrey Gordon, and Alex Smola. Hilbert Space Embeddings of Hidden Markov Models. In *The 27th International Conference on Machine Learning (ICML2010)*, pages 991–998, 2010.

Zhenyue Zhang and Jing Wang. MLLE: Modified Locally Linear Embedding Using Multiple Weights. *Advances in Neural Information Processing Systems*, pages 1593–1600, 2006. ISSN 10495258.