# Empowering Variational Inference with Predictive Features: Application to Disease Suptypying


## Abstract

Generative models, such as the probabilistic topic model, have been widely deployed for various applications in the healthcare domain, such as learning disease or tissue subtypes. However, learning the parameters of such models is usually an ill-posed problem and may lose valuable information about disease severity. A common approach is to add a discriminative loss to the generative model's learning loss; finding a balance between two losses is not straightforward. We propose an alternative way in this paper. We use distribution embedding to construct patient-level representation, which is usually more discriminative than the posterior parameters. We view the patient-level representation as an external covariate. Then, we use the external covariates to inform the posterior of our generative model. Effectively, we enforce the generative model's approximate posterior to reside in the subspace of the discriminative covariates. We illustrate this 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 a topic model as a generative model. We evaluate the patient representation, the resulting topics on the patient- and population-level. We also show that some of the discovered subtypes are correlated with genetic measurements, suggesting that the identified subtypes characterize the disease's underlying etiology.


## 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 flexibility 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 different 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 model to formulate the subtype discovery problem.

We view 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." Such a specific 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 patient representation. The 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, hence indirectly encourage subtypes to be disease-related. Instead of supervised topic modeling, we propose to inject discriminative information in the form of covariates to the subtypes' inference model (*i.e.,* topics).

**Related Works.** Various unsupervised subtype discovery methods have been proposed. 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 sub-populations 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 inject the 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:

- 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.,* mixture models (Hannah et al., 2011) or hidden Markov models (Moscovich and Chen, 2004)).

- 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 model 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).

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

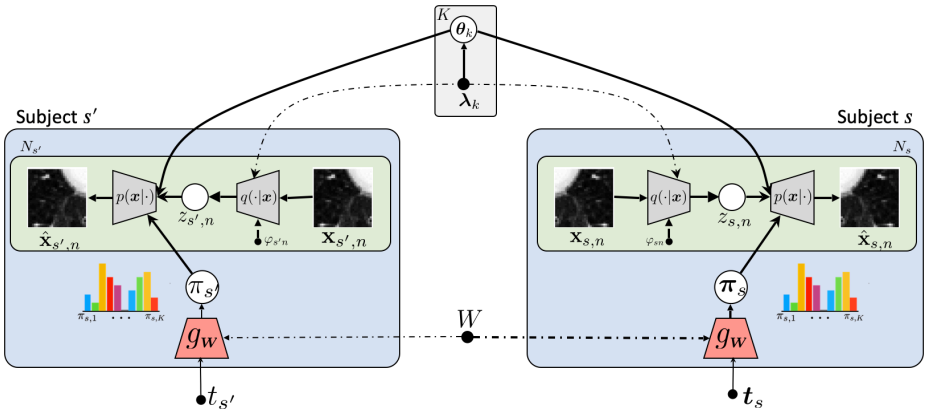



Figure 1: The schematic of our framework for two subjects $s$ and $s'$ with $t_s$ and $t_{s'}$ as their corresponding covariates. The encoder ($q(\cdot|x)$) and decoder ($p(x|\cdot)$) inside the green box explain data at the supervoxel-level (word-level) while the $g_w$ explains the subject-level data (*i.e.,* topic proportion). The $\theta_k$ and $\lambda_k$ 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.

## 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 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

| | |
|---|---|
| **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 $\boldsymbol{\pi}_s$. |
| $\alpha$ | 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 $\boldsymbol{\theta}_k$. |
| $\beta^*$ | 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.

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 $\boldsymbol{\Sigma}_k$,

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

where $\text{NIW}(\eta)$ is the Normal-Inverse-Wishart distribution with hyper-parameter $\eta$. 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:

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

where the $\boldsymbol{\pi}_s$ follows the Dirichlet distribution, $\text{Cat}(\boldsymbol{\pi}_s)$ 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 $\beta_k$'s are concentration parameters. If $\beta_k$'s are greater than one, the topics distribution becomes more disperse (less sparse).

To avoid tuning $K$ hyper-parameters for $\beta_1$ to $\beta_K$, we follow the truncated Hierarchical Dirichlet Process (HDP) (Teh et al., 2006), and assume $\beta$ is generated by the so-called "stick-breaking" construction,

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

where $\text{Beta}(\cdot, \cdot)$ indicates the Beta distribution. Such construction allows for controlling the sparseness of the topics distribution with a single hyper-parameter (*i.e.,* $\alpha$) 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,

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

### 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 $\beta^*$, $\boldsymbol{\varphi}_{sn}$, $\boldsymbol{\lambda}_k$, and $\boldsymbol{\omega}_s$ are the variational parameters corresponding to the random variables $\beta$, $z_{sn}$, $\boldsymbol{\theta}_k$, 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 $z_{s,\cdot}$ 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 the 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.

Unlike the rest of the variables, $\boldsymbol{\pi}_s$ is defined at the *subject-level,* characterizing the topics proportion for subject $s$. The $\boldsymbol{t}_s$ is also 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 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 the 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:

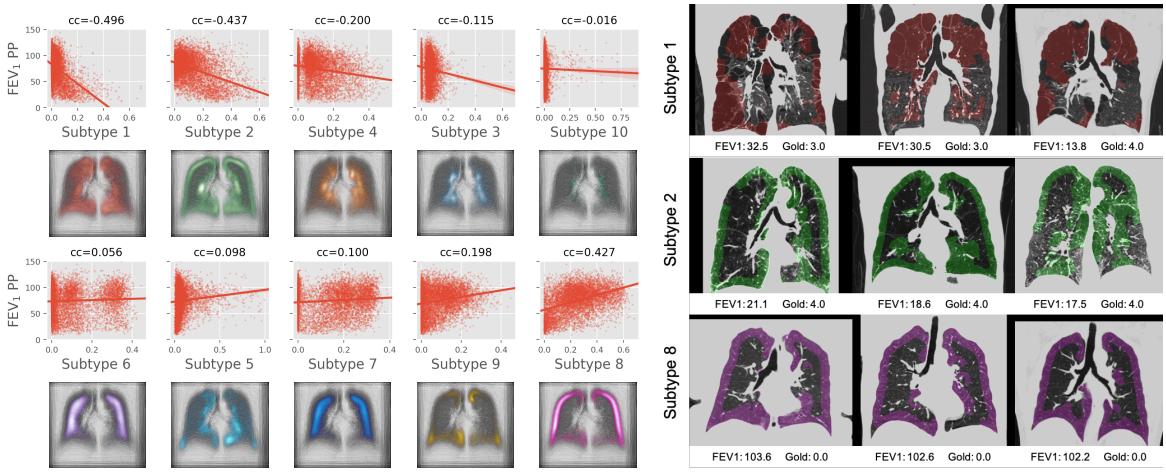
$$\begin{aligned} \boldsymbol{\epsilon} &\sim \mathcal{N}(0, I_{K\times K}) \\ \boldsymbol{\psi}_s &= \boldsymbol{\mu}(\boldsymbol{t}_s; \boldsymbol{W}_{\boldsymbol{\mu}}) + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}(\boldsymbol{t}_s; \boldsymbol{W}_{\boldsymbol{\sigma}}) \\ \boldsymbol{\pi}_s &=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

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.,* $h_{SB} : \mathbb{R}^K \to \Delta^K$. 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,

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




(a) (b)

Figure 2: (a) *Odd Rows:* Pearson correlation between proportion of subtype and $\text{FEV}_1$. The $x$- and $y$-axis are the subtype proportion and $\text{FEV}_1$ 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 $\text{FEV}_1$), whereas *subtype 8* appears to be healthy tissue (positively correlated with $\text{FEV}_1$).

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 without discriminative features injection and a supervised topic model.

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 oversegment 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 details 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 derived from two Spirometry measurements and is between zero and four where zero is used for people at risk (Normal Spirometry but Chronic Symptoms), 1-4 denote Mild to Very Severe COPD. The -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 $\text{FEV}_1$ PP and $\text{FEV}_1/\text{FVC}$ from the subject-level features ($\boldsymbol{t}_s$). We use the predicted values to compute the GOLD score[1].

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 $\pi_s$. More concretely, we assume $y_s \sim \mathcal{N}(\boldsymbol{\mu}(\pi_s), \sigma^2)$ where $\boldsymbol{\mu}$ is a learnable function and $\sigma^2$ 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.

1. 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.

| | $R^2$ | | | |
|---|---|---|---|---|
| Subject-Level Descriptor | $FEV_1$ PP | $FEV_1$/FVC | FVC | Distance Walked |
| %Low Attenuation Level (-950) | 0.44 | 0.61 | 0.03 | 0.07 |
| 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 |

Table 3: Performance of predicting $FEV_1$ PP, $FEV_1$/FVC, FVC, and distance walked compared across 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 rest 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 $\pi_s$ of subject $s$.

**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* injected 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.

### 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 Figure 2($b$)subfigure for a few subjects and subtypes. We register the label masks of all the subtypes to a common space to compute the average distribution of each subtype across the population. Figure 2($a$)subfigure 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

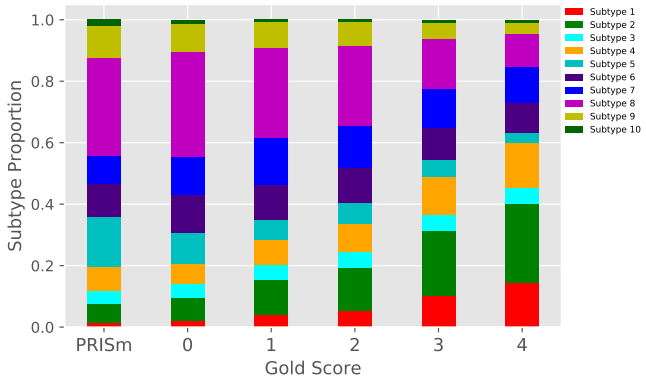



Figure 3: Subtype proportions averaged over subsets of the population with GOLD score values PRISm, 0, 1, 2, 3, and 4.

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 Figure 3. Each bar represents a sub-population of patients with a particular GOLD score and colors within the bar represent 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 $\text{FEV}_1$. Figure 2($b$)subfigure 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 $\geq 3$). 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 $\text{FEV}_1$ 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.

| 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 | $\approx 0$ | 8.17 | 5e-01 |
| 6 | $\approx 0$ | 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.

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 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 so-called 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 $\text{FEV}_1$, *subtype 8* has the strongest positive correlation with the $\text{FEV}_1$.

## 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 inject $\boldsymbol{t}_s$ 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 validated the main idea of the paper by empowering our variational posterior distribution with these predictive features. 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. Figures 3.1 visualizes 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 $\text{FEV}_1$ 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 (Figure 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 $\text{FEV}_1$. Understanding the biological etiology of those 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.

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.

LG Moscovich and Jianhua Chen. Supervised hidden markov model learning using the state distribution oracle. In *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, volume 1, pages 240–244. IEEE, 2004.

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.

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.

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.

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.

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.