Unsupervised Discovery of Emphysema Subtypes in a Large Clinical Cohort

Polina Binder[1](✉), Nematollah K. Batmanghelich[2], Raul San Jose Estepar[2], and Polina Golland[1]

[1] Computer Science and Artificial Intelligence Lab, EECS, MIT, Cambridge, USA
polinab@mit.edu

[2] Brigham and Womens Hospital, Harvard Medical School, Boston, USA

Abstract. Emphysema is one of the hallmarks of Chronic Obstructive Pulmonary Disorder (COPD), a devastating lung disease often caused by smoking. Emphysema appears on Computed Tomography (CT) scans as a variety of textures that correlate with disease subtypes. It has been shown that the disease subtypes and textures are linked to physiological indicators and prognosis, although neither is well characterized clinically. Most previous computational approaches to modeling emphysema imaging data have focused on supervised classification of lung textures in patches of CT scans. In this work, we describe a generative model that jointly captures heterogeneity of disease subtypes and of the patient population. We also describe a corresponding inference algorithm that simultaneously discovers disease subtypes and population structure in an unsupervised manner. This approach enables us to create image-based descriptors of emphysema beyond those that can be identified through manual labeling of currently defined phenotypes. By applying the resulting algorithm to a large data set, we identify groups of patients and disease subtypes that correlate with distinct physiological indicators.

1 Introduction

Chronic Obstructive Pulmonary Disorder (COPD) is a chronic lung disease characterized by poor airflow. One of the hallmarks of COPD is emphysema, i.e., destruction of lung alveoli and permanent enlargement of airspaces [1]. Several subtypes of emphysema have been identified and are commonly used for diagnosis and prediction of patient prognosis [2]. The disease subtypes have also been shown to correlate with genetic data and physiological indicators [1].

Emphysema appears on Computed Tomography (CT) scans as a variety of textures which are associated with clinically defined disease subtypes. However, there is substantial intra-reader and inter-reader variability when identifying subtypes in CT images [2]. Computational approaches to the classification of textures in CT scans promise to identify subtle textural differences beyond those that are visible to human readers. This nuanced information can be harnessed to produce well-defined, reproducible disease subtypes. Beyond fully 3D texture

L. Wang et al. (Eds.): MLMI 2016, LNCS 10019, pp. 180–187, 2016.
DOI: 10.1007/978-3-319-47157-0_22

analysis, the additional benefits of computational approaches include the possibility of providing novel insights into the disease once the heterogeneity of the patient population is characterized.

We present a method that simultaneously detects distinct patient clusters and disease subtypes. The algorithm is based on a generative model that captures the underlying hypothesis about population structure and distributions of disease subtypes. We assume that each cluster of patients is associated with a distinct distribution of disease subtypes, which are based on features extracted from Computed Tomography scans [3]. We derive an inference algorithm that is based on variational Expectation-Maximization [4]. We apply the algorithm to a data set of 2457 thoracic CT scans and observe notable associations between physiological indicators and patient clusters and disease subtypes identified by the method. Further, we examine associations in simplified models that omit either patient clusters or disease subtypes to demonstrate the clinical advantage of the hierarchical model that includes both patient clusters and disease subtypes. We compare associations that are identified in the generative model to those found in a model where disease subtypes are discovered in a supervised manner.

Our approach departs from the majority of prior research that has focused on supervised classification of patches extracted from CT scans based on examples labeled by clinical experts [5,6]. An exception is a method for joint modeling of imaging and genetic data in the same clinical population [7]. By contrast, our work models only imaging data, but we explicitly detect and characterize homogeneous sub-populations defined by similar groups of disease subtypes, which opens directions for future analysis. An additional work similar to ours is found in [8], which discovers disease subtypes in an unsupervised manner. However, it was conducted on a smaller data set and does not model patient clusters.

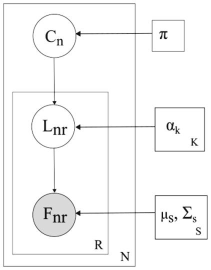

Fig. 1. Graphical representation of the generative model.

2 Model

Our generative model relies on the assumption that there are K underlying patient clusters, each characterized by a different distribution of disease subtypes. We use N to denote the total number of CT scans in the study. When processed, each scan is represented by R non-overlapping patches. Let S_{nr} be the patch around voxel r in patient n. Patches are entirely contained within a lung. We apply a chosen feature extraction method to S_{nr} to construct a feature vector F_{nr}. The feature vectors $\{F_{nr}\}$ serve as the input into our algorithm. In our experiments we use a combination of Grey Level Co-Occurrence Matrix (GLCM) [6] features and intensity histograms as feature descriptors which are both extracted from three-dimensional patches; the modeling approach readily accepts a broad range of descriptors.

The distribution of cluster assignments for any patient in the study is parametrized by π and is represented by a vector C_n for patient n. $C_{nk} = 1$ if patient n belongs to cluster k; $C_{nk} = 0$ otherwise. For all patients in cluster k the distribution of disease subtypes is parametrized by α_k and is represented by L_{nr} for patch r in patient n. Each patch belongs to one of S disease subtypes. $L_{nrs} = 1$ if the patch belongs to subtype s; $L_{nrs} = 0$ otherwise. We use a Gaussian distribution $\mathcal{N}(\cdot; \mu, \Sigma)$ with mean μ_s and covariance Σ_s to model feature vectors in the disease subtype s. The generative model can be summarized as follows (Fig. 1):

$$C_n \sim \prod_{k=1}^{K} \pi_k^{C_{nk}},$$

$$L_n | C_n \sim \prod_{k=1}^{K} \prod_{r=1}^{R} \prod_{s=1}^{S} (\alpha_{ks})^{L_{nrs} C_{nk}},$$

$$F_n | L_n \sim \prod_{s=1}^{R} \prod_{r=1}^{S} \mathcal{N}(F_{nr}; \mu_s, \Sigma_s)^{L_{nrs}}.$$

Each subject is viewed as an independent and identically distributed sample from this distribution, giving rise to the full likelihood model:

$$p(F, C, L; \alpha, \pi, \mu, \Sigma) = \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{r=1}^{R} \prod_{s=1}^{S} \left(\pi_k {\alpha_{ks}}^{L_{nrs}}\right)^{C_{nk}} \mathcal{N}(F_{nr}; \mu_s, \Sigma_s)^{L_{nrs}}.$$

Inference Algorithm. We set the number of patient clusters K and the number of disease subtypes S. The observed data consists of feature vectors $\{F_{nr}\}$ of N patients for whom we extracted features from R patches each. We aim to infer the most likely subtype L_{nr} for each patch r in patient n and the most likely cluster C_n for each patient n. Additionally, we estimate the parameters: the mixing proportions of the patient clusters π, the mixing proportions of the disease subtypes $\{\alpha_k\}$ for each patient cluster, and the means and variances $\{\mu_s, \Sigma_s\}$ of the image features for each disease subtype.

We perform inference via variational Expectation-Maximization (EM) [4]. Since computing expectation with respect to the full posterior distribution $p(L, C | F, \alpha, \pi, \mu, \Sigma)$ is intractable due to coupling between C and L, we approximate the posterior distribution with a product of two categorical distributions:

$$q(C, L; \psi, \theta) = q_C(C; \psi) q_L(L; \theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \psi_{nk}^{C_{nk}} \prod_{r=1}^{R} \prod_{s=1}^{S} \theta_{nrs}^{L_{nrs}}, \tag{1}$$

where ψ and θ are variational parameters. This simplifies the computation of the expectations.

In the variational approach, we iteratively optimize a lower bound for $\ln(p(F; \alpha, \pi, \mu, \Sigma))$ with respect to the parameters $\{\pi_k, \alpha_{ks}, \mu_s, \Sigma_s, \psi_{nk}, \theta_{nrs}\}$, This lower bound can be expressed as:

$$\ln(p(F; \alpha, \pi, \mu, \Sigma)) \geq E_q \left[\ln \frac{p(F, C, L; \alpha, \pi, \mu, \Sigma)}{q(C, L; \psi, \theta)} \right]. \tag{2}$$

We randomly initialize π and α, and then iterate between two steps until convergence. In the expectation step, we hold π, α, μ and Σ fixed and estimate the variational parameters ψ and θ to maximize the lower bound in Eq. (2) by iteratively applying the updates:

$$\psi_{nk} \propto \prod_{s=1}^{S} \prod_{r=1}^{R} \alpha_{ks}^{\theta_{nrs}}, \quad \text{s.t.} \quad \sum_{k=1}^{K} \psi_{nk} = 1,$$

$$\theta_{nrs} \propto \prod_{k=1}^{K} \alpha_{ks}^{\psi_{nk}}, \quad \text{s.t.} \quad \sum_{s=1}^{S} \theta_{nrs} = 1.$$

In the maximization step, we hold the values of ψ and θ fixed and estimate the model parameters, π, α, μ and Σ, that maximize the lower bound in Eq. (2) via the following update equations:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \psi_{nk},$$

$$\alpha_{ks} \propto \sum_{n=1}^{N} \psi_{nk} \sum_{r=1}^{R} \theta_{nrs}, \quad \text{s.t.} \quad \sum_{s=1}^{S} \alpha_{ks} = 1,$$

$$\mu_s = \frac{1}{N_s} \sum_{n=1}^{N} \sum_{r=1}^{R} \theta_{nsr} \cdot F_{nr}, \quad \text{where} \quad N_s = \sum_{n=1}^{N} \sum_{r=1}^{R} \theta_{nsr},$$

$$\Sigma_s = \frac{1}{N_s} \sum_{n=1}^{N} \sum_{r=1}^{R} \theta_{nsr} \cdot (F_{nr} - \mu_s) \cdot (F_{nr} - \mu_s)^T.$$

Once the parameter estimation process is complete, we determine C_n and L_{nr} by maximizing the approximate posterior distributions $q_C(C_n; \psi_n)$ and $q_L(L_{nr}; \theta_{nr})$ respectively.

3 Empirical Results

Data. We investigated the proposed method in the context of an imaging study that includes 2457 thoracic CT scans of smokers diagnosed with COPD [1]. COPDGene is a multi-center study that acquired CT scans, genetic data, and physiological indicators in COPD patients. The data was collected by 21 sites across the United States. The volumetric CT scans were obtained at full inhalation and at relaxed exhalation. Image reconstruction produces sub-millimeter slice thickness, and employs edge and smoothness enhancing filtering [1]. In addition, we have 1525 patches from the CT scans of 267 patients from this cohort that were manually assigned to clinically defined disease subtypes by an expert.

Parameter Selection. We randomly sampled 1000 non-overlapping patches from each patient. Emphysema has been described at the level of the secondary pulmonary lobules [5], therefore we select $11 \times 11 \times 11$ patches, which are approximately the size of this structure. There have been between four and 12 disease subtypes and between three and 10 patient clusters described in clinical literature [3,5]. We examine models with the number of patient clusters and disease subtypes in this range. We chose to further analyze the model with eight patient clusters and six disease subtypes, as this was the largest number of disease subtypes and patient clusters for which each patient cluster and disease subtype received at least five percent probability.

Feature Vectors. We employed 11-dimensional feature vectors, which were chosen based on their classification accuracy on the labeled patches in our data set when using the features as a texture descriptor. The first nine dimensions correspond to Grey Level Co-Occurrence Matrix (GLCM) features [6]. GLCMs represent the joint probability distribution of intensity values of pixel pairs in a given patch [6]. To construct this descriptor, the image is discretized into eight gray levels. The value of the entry at position (i, j) in the GLCM captures the proportion of pixel pairs at a given offset with the corresponding intensity pair values for $i, j \in \{1...8\}$. To obtain a degree of rotational invariance, we averaged the GLCMs over uniformly distributed directions in three dimensions. We extracted nine features from these matrices to construct the descriptor: contrast, dissimilarity, homogeneity, correlation, entropy, energy, cluster shade, cluster prominence and maximum probability [6]. The next two dimensions of the feature vector correspond histogram bins of the voxel intensities within the patch.

3.1 Results

Disease Subtypes. Figure 2 illustrates example patches for each of the identified disease subtypes. A confusion matrix between the disease subtypes and the clinical labels is shown in Table 1. On the labeled portion of our data set, we found that 67 % of patches that were labeled as clinically normal were placed in the same disease subtype by our algorithm, and clinically normal patches represent 64 % of all labeled patches within this disease subtype. Panlobular and paraseptal emphysema correspond to disease subtype 2 and subtype 3 respectively. Our results suggest that centrilobular emphysema is a mixture of identified disease subtypes 1, 2, 3 and 4.

Table 1. Confusion matrix between clinically defined subtypes and automatically detected subtypes. The values in the table correspond to the number of patches with the corresponding clinical label and detected subtype.

Clinical label	ST 1	ST 2	ST 3	ST 4	ST 5	ST 6
Normal lung tissue	339	0	1	103	7	61
Panlobular emph.	1	146	9	0	0	0
Paraseptal emph.	16	53	100	48	20	6
Mild centrilobular emph.	96	3	11	68	3	30
Mod. centrilobular emph.	69	74	112	28	4	2
Sev. centrilobular emph.	8	57	49	0	0	0

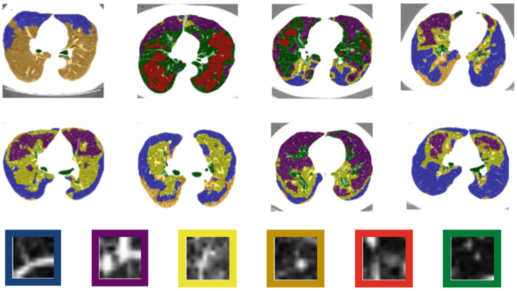

Fig. 2. Top two rows: example CT scans from each of the eight patient clusters identified by our algorithm. Colors correspond to disease subtypes identified by our algorithm. Bottom row: patches from the six disease subtypes identified by our algorithm. (Color figure online)

Spatial Contiguity. Emphysema clusters spatially in the lungs, as do the disease subtypes our algorithm identifies, as can be seen in Fig. 2. Each voxel in every lung was labeled independently based on the most likely subtype it would belong to under our model, without any enforced smoothing. We evaluated spatial contiguity by permutation testing [9]. For each voxel labeled by our algorithm we compute the proportion of neighboring voxels that belong to the same disease subtype. We average this value over the entire lung to obtain a spatial contiguity score. To obtain a distribution of the score under the null hypothesis we assigned voxels within the lungs to random disease subtypes 1000 times for each scan while maintaining the proportion of disease subtypes for each lung. We found that across all CT scans, the spatial contiguity scores produced by our algorithm are greater than the maximal values in the corresponding null distribution, corresponding to rejecting the null hypothesis with $p < 0.001$.

Associations with Physiological Indicators. We emphasize that the physiological indicators are not available to the algorithm when fitting the generative model to the image data and therefore provide an indirect validation of the model's clinical relevance. We quantify the associations between the structure detected by our method and physiological indicators relevant to COPD: six minute walking distance, body mass index (BMI), forced vital capacity (FVC), forced expiratory volume (FEV), change in FVC value from treatment, the ratio between the FEV and FVC values, and the number of years smoked. We ran our algorithm on a randomly selected half of our scans and labeled the remaining scans based on the estimated model parameters. In particular, we assigned each patient to the most likely cluster and constructed an empirical distribution of disease subtypes for the patient based on the image patches. We repeated this procedure 100 times to estimate variability in the results.

We constructed three baseline models by eliminating patient clusters ($K = 1$) or disease subtypes ($S = 1$) or both ($K = 1, S = 1$). In the last case, we extract feature vectors from patches in each patient, and then average and normalize the feature vectors in each patient to produce a single patient-specific feature vector. A fourth baseline method was constructed by identifying the disease subtypes in a supervised manner. In this case, we utilized the same feature vectors as previously described, and performed classification with Support Vector Machines (SVMs) trained on the labeled patches to assign 1000 random patches in each lung to one of six clinically identified subtypes. We learned the patient clusters in an unsupervised manner as in the fully unsupervised model.

To quantify the associations between distributions of disease subtypes or the averaged normalized feature vector for a patient and a physiological indicator we perform linear regression. The strength of the correlation is quantified via the R^2 value. The association between patient clusters and physiological indicators is quantified via the normalized mutual information score [10]. Different metrics are used to quantify the associations between patient clusters and proportions of disease subtypes or feature vectors, as the former is a discrete label while the last two are continuous quantities. These associations were identified on the portion of the data set that was not used to construct the model.

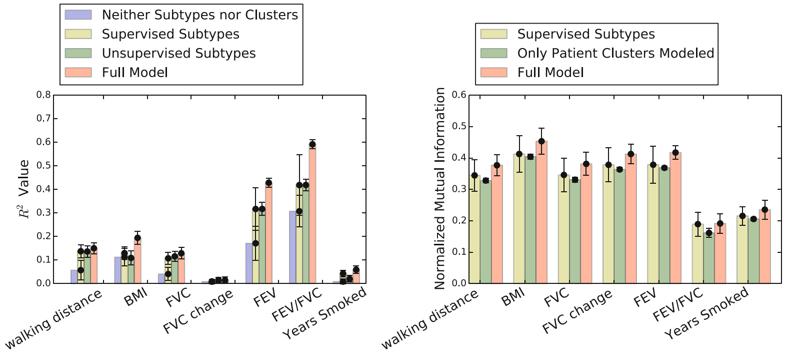

Fig. 3. Left: R^2 value between the distributions of disease subtypes (1st, 3rd, and 4th model) or feature vectors (2nd model) and physiological indicators. Right: Normalized Mutual Information between patient clusters and physiological indicators.

Figure 3 reports the associations for all models. These results demonstrate the advantage of modeling both patient clusters and disease subtypes. We observe that there is a stronger association between physiological indicators and patient clusters in the full model than in the model with only clusters. For all physiological indicators, there is a higher association with the distributions of disease subtypes in the full model than in the model with only disease subtypes. This demonstrates that modeling patient clusters produces more clinically relevant distributions of disease subtypes in each patient. The model without patient clusters or disease subtypes exhibits even weaker associations than a model with only disease subtypes.

Figure 3 demonstrates the advantage of discovering the disease subtypes in an unsupervised manner. In the full model, we obtain stronger associations than in the model where disease subtypes are found in a supervised manner. This is partially explained by the fact that feature selection was performed to optimize classification performance on supervised patches and additional structure is obtained in the unsupervised discovery of patient clusters and disease subtypes.

4 Conclusions

We presented an unsupervised framework for the discovery of disease subtypes within emphysema and of patient clusters that are characterized by distinct distributions of such subtypes. We built a generative model that parametrizes the assignment of voxels in CT scans to disease subtypes and the assignment of patients to clusters. The associations between the patient clusters and physiological indicators and distributions of disease subtypes and physiological indicators illustrate the clinical relevance of the detected heterogeneity in the patient cohort.

The patient clusters that our model produces merit further exploration. It would be worthwhile to examine their correlations to genetic markers. An additional extension is to directly examine whether different patient clusters exhibit distinct clinical prognoses or respond differently to clinical interventions.

References

1. Regan, E.A., Hokanson, J.E., et al.: Genetic epidemiology of COPD (COPDGene) study design. COPD **7**(1), 32–43 (2010)
2. Aziz, Z., Wells, A., et al.: HRCT diagnosis of diffuse parenchymal lung disease: inter-observer variation. Thorax **59**(6), 506–511 (2004)
3. Raghunath, S., Rajagopalan, S., et al.: Quantitative stratification of diffuse parenchymal lung diseases. PloS ONE **9**, e93229 (2014)
4. Blaiotta, C., Cardoso, M.J., et al.: Variational inference for image segmentation. Comput. Vis. Image Underst. (2016)
5. Mendoza, C., Washko, G., et al.: Emphysema quantification in a multi-scanner HRCT cohort using local intensity distributions. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 474–477 (2012)
6. Prasad, M., Sowmya, A., et al.: Multi-level classification of emphysema in HRCT lung images. Pattern Anal. Appl. **11**(1), 9–20 (2006)
7. Batmanghelich, N.K., Saeedi, A., Cho, M., Estepar, R.S.J., Golland, P.: Generative method to discover genetically driven image biomarkers. In: Ourselin, S., Alexander, D.C., Westin, C.-F., Cardoso, M.J. (eds.) IPMI 2015. LNCS, vol. 9123, pp. 30–42. Springer, Heidelberg (2015). doi:10.1007/978-3-319-19992-4_3
8. Hame, Y., Angelini, E.D., et al.: 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), pp. 109–113. IEEE (2015)
9. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)
10. Vinh, N.X., Epps, J., et al.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2537–2854 (2010)