# Manifold Learning of COPD

Felix J.S. Bragman[1], Jamie R. McClelland[1], Joseph Jacob[1]
John R. Hurst[2] and David J. Hawkes[1]

[1] Centre for Medical Image Computing, University College London, UK
[2] UCL Respiratory, University College London, UK

**Abstract.** Analysis of CT scans for studying Chronic Obstructive Pulmonary Disease (COPD) is generally limited to mean scores of disease extent. However, the evolution of local pulmonary damage may vary between patients with discordant effects on lung physiology. This limits the explanatory power of mean values in clinical studies. We present local disease and deformation distributions to address this limitation. The disease distribution aims to quantify two aspects of parenchymal damage: locally diffuse/dense disease and global homogeneity/heterogeneity. The deformation distribution links parenchymal damage to local volume change. These distributions are exploited to quantify inter-patient differences. We used manifold learning to model variations of these distributions in 743 patients from the COPDGene study. We applied manifold fusion to combine distinct aspects of COPD into a single model. We demonstrated the utility of the distributions by comparing associations between learned embeddings and measures of severity. We also illustrated the potential to identify trajectories of disease progression in a manifold space of COPD.

## 1 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a complex disorder arising from various pathological processes including emphysema and functional small airways disease (fSAD). The extent of emphysema and fSAD that make up overall disease burden can vary, which can affect lung physiology. Both disease processes can progress at different rates, complicating prognostication. Optimising the quantification of disease extent in COPD may improve the precision of disease staging and monitoring.

Analysis of lung disease from Computed Tomography (CT) has typically relied on the analysis of the lung using global averages. Such metrics cannot capture the anatomical distribution of disease. Methods have been proposed to quantify the contribution of various emphysema subtypes [5] or the distribution of image features [2]. Harmouche et al. [5] built an emphysema manifold by analysis of classified emphysema subtypes. A Severity Index (S) was derived from this space that is complimentary to the mean level of emphysema. In contrast, Bragman et al. [2] modelled local distributions of density and biomechanical features; exploiting them to investigate differences between subtypes of COPD whilst also classifying these subtypes.

## 2 Method

We present a new method to quantify the spread of parenchymal disease and measure its effect on lung deformation. It is based on locally quantifying tissue destruction and deformation to capture heterogeneity or homogeneity across the lung. The outcome is a distribution that quantifies various aspects of lung pathophysiology that can be modelled to test associations with various clinical hypotheses. The distributions can be exploited to quantify inter-patient differences in lung tissue pathology and deformation. A single model of tissue disease and deformation can be obtained by combining separate embeddings obtained from manifold learning with manifold fusion.

### 2.1 Lung deformation and tissue classification

The deformation between paired breath-hold CT scans acquired at forced residual capacity ($\mathcal{I}_{exp}$, $\Omega^*$) and total lung capacity ($\mathcal{I}_{ins}$, $\Omega$) can be obtained using nonrigid registration. The output is a transformation $\varphi$ mapping each coordinate $x \in \Omega \rightarrow x^* \in \Omega^*$. Local volume change is characterised by the Jacobian determinant $J$. It is calculated on a voxel-wise basis: $J = \det(\nabla_x \varphi)$.

Parametric Response Mapping (PRM) [4] was used to classify voxels as emphysema (PRM$_{emph}$) and functional small airways disease (PRM$_{fSAD}$). For all voxels $x_i \in \mathcal{I}_{ins}$, the tissue class $z_i$ is based on Hounsfield Unit (HU) thresholds in $\mathcal{I}_{ins}$ and $\mathcal{I}_{exp}$. A voxel is classified as PRM$_{emph}$ if $\mathcal{I}_{ins}(x_i) \leq -950$ and $\mathcal{I}_{exp}(\varphi(x_i)) \leq -856$. A voxel is classified as PRM$_{fSAD}$ if $\mathcal{I}_{ins}(x_i) > -950$ and $\mathcal{I}_{exp}(\varphi(x_i)) \leq -856$. The airways and vasculature are segmented by only considering voxels with an HU between $-500$HU and $-1024$HU in both scans.

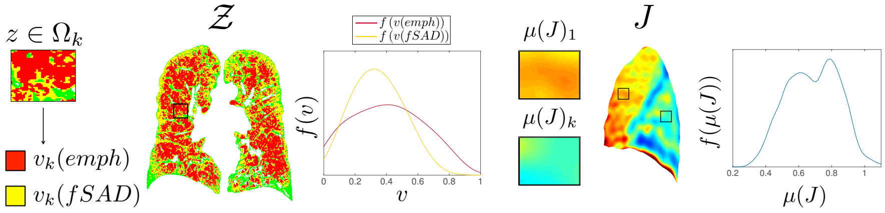
### 2.2 Local disease and deformation distributions

We present the concept of local feature distributions (Fig.1a and b). The aim is to quantify local abnormalities in lung physiology and pathology to define a signature unique to a patients disease state. We introduce two models: 1) local disease distributions and 2) local deformation distributions. The disease distributions model the spread of emphysema and fSAD whilst the deformation distribution characterises local volume change across the lung. They are created by locally sampling regions of $\mathcal{Z}$ and $J$ in a Cartesian grid using local regions of interest $\Omega_k$ (ROI) where $k = 1 \cdots K$ indexes the center voxel of the ROI. The size $(r \times r \times r)$ of the ROI governs the scale of the sampling.

We modelled two properties of disease spread: 1) locally diffuse/dense disease and 2) global homogeneity/heterogeneity. For each ROI centered at $z_k$ where $z \in \Omega_k$, we computed the fraction of PRM$_{emph}$ and PRM$_{fSAD}$ voxels; defined as $v_k(emph)$ and $v_k(fSAD)$. Dense disease occurred when $v_k(\cdot) \rightarrow 1$ whilst diffuse disease was present when $v_k(\cdot) \rightarrow 0$. The deviation of diffuse and dense regions in the lung defined the heterogeneity/homogeneity of disease spread.

A distribution $f(v(\cdot))$ for each feature was built by sampling $K$ regions. The shape of the distribution is governed by the two disease properties (Fig.1a). It provides information on the nature of local disease spread (diffuse or dense) and whether it is homogeneous or heterogeneous.

Expansion of the lung is dependent on local biomechanical properties (emphysema) and airway resistance (functional small airways disease), which will affect lung deformation locally. To capture volume change on a local basis, the Jacobian map ($J$) was sampled by calculating the mean Jacobian ($\mu(J)_k$) for all $\Omega_k$. A distribution $f(\mu(J))$ of these measurements was built to capture local volume change throughout the lung using the same process as above (Fig.1b).



(a) Local disease

(b) Local deformation

**Fig. 1.** Local disease and deformation distributions.

### 2.3 Manifold learning of COPD distributions

We hypothesised that the heterogeneity of COPD could be modelled by the local disease and deformation distributions. Manifold learning can be used to capture variability in the distributions and learn separate embeddings for emphysema, fSAD and lung deformation. Fusion of these embeddings can then be performed to create various models of COPD.

**Distribution distance.** Inter-patient differences are computed using the Earth Movers Distance ($\mathcal{L}_{EMD}$) [11]. It is a cross-bin distance metric, which measures the minimum amount of work needed to transform one distribution into another. The distributions are quantised into separate histograms $h_{v(emph)}$, $h_{v(fSAD)}$ and $h_J$ using $N_b$ bins. They are normalised to sum to 1 such that they have equal mass. A closed-form solution of the $\mathcal{L}_{EMD}$ can be used for one-dimensional distributions with equal mass and bins [7]. It reduces to the $\mathcal{L}_1$-norm between cumulative distributions ($H$) of two histograms $h_{1,(\cdot)}$ and $h_{2,(\cdot)}$: $\mathcal{L}_{EMD}\left(h_{1,(\cdot)}, h_{2,(\cdot)}\right) = \left(\sum_n^{N_b} |H_{n,1,(\cdot)} - H_{n,2,(\cdot)}|\right)$.

**Manifold learning and fusion.** Manifold learning is used to model emphysema, fSAD and Jacobian distributions. The aim is to capture variations in the distributions in a population of COPD patients. As emphysema and fSAD occur synchronously and both affect lung function, the manifold fusion framework of Aljabar et al. [1] is employed to create a single representation of these processes.

For $P$ subjects, the PRM classified volumes are $\mathcal{Z}_1, \cdots, \mathcal{Z}_P$ and their respective Jacobian determinant maps are $J = J_1, \cdots, J_P$. The distributions are quantised using $N_b$ bins into their respective histograms $h_{p,v(emph)}$, $h_{p,v(fSAD)}$ and $h_{p,J}$. Pairwise measures in the population are obtained with the $\mathcal{L}_{EMD}$ yielding the pairwise matrices $\mathcal{M}^{emph}$, $\mathcal{M}^{fSAD}$ and $\mathcal{M}^J$. They can be visualised as connected graphs where each node represents a patient and the edge length is the $\mathcal{L}_{EMD}$. Isomap[1] [12] is applied to each matrix. A $K$-nearest neighbour search is first performed to create a sparse representation of $\mathcal{M}^{(\cdot)}$ where edges are restricted to the $K$-nearest neighbourhood of each node. A full pairwise geodesic distance matrix $D^{(\cdot)}$ is then estimated by analysis of the $K$-nearest graph of $\mathcal{M}^{(\cdot)}$ using Djikstra's shortest-path algorithm [3]. The low-dimensional embedding $y_p^{(\cdot)}, p = 1, \cdot, P$ is obtained by minimisation of

$$\min \sum_{p,j} \left( D_{p,j}^{(\cdot)} - ||y_p^{(\cdot)} - y_j^{(\cdot)}|| \right)^2 \tag{1}$$

using Multi-Dimensional Scaling. The coordinate embeddings for $\mathcal{M}^{emph}$, $\mathcal{M}^{fSAD}$ and $\mathcal{M}^J$ are $y^e$, $y^f$ and $y^J$ with dimensions $d^e$, $d^f$ and $d^J$ that are selected.

Fusion of the coordinates $y^{(\cdot)}$ can be performed in any combination to investigate various processes. For simplicity, we consider all embeddings. The coordinates are uniformly scaled with the scale factors $s^e$, $s^f$ and $s^J$ such that the first component of each embedding $y_1^{(\cdot)}$ has a unit variance. These are concatenated to yield $Y = (s^e y^e, s^f y^f, s^J y^J)$ with dimension $d^e + d^f + d^J$. A distance matrix $\mathcal{M}^c$ is obtained by calculating pairwise Euclidean distances of $Y$. Isomap is then applied to yield the combined coordinate embedding $y^c$ with dimension $d^c$.

## 3 Experiments

### 3.1 Data processing

A total of $1,154$ scans of COPD patients (GOLD $\geq 1$) were downloaded from COPDGene [10]. They were acquired on various scanners (GE Medical Systems, Siemens and Philips) with the following reconstruction algorithms: STANDARD (GE), AS+ B31f and B31f (Siemens), and 64 B (Philips). The Pulmonary Toolkit[2] was used for lung segmentation. Breath-hold scans were registered with NiftyReg [9] with a modified version of the EMPIRE10 pipeline [8]. The transformation was a stationary velocity field parameterised by a cubic B-spline and the similarity measure was MIND [6]. The constraint term was the bending energy of the velocity field, weighted at 1% for all stages of the pipeline. After manual inspection of the registrations, 743 patients were selected. Scans were rejected if there were major errors close to the fissures and the lung boundary.

The sampling size of the ROIs was $r = 20$mm, consistent with the size of the secondary pulmonary lobule. Sampling was performed with a Cartesian grid of

[1] lvdmaaten.github.io/drtoolbox/

[2] github.com/tomdoel/pulmonarytoolkit

center voxels spaced every 5mm. We chose a value of $N_b = 60$ as its effect on pairwise distances was minimal with increasing $N_b$ when $N_b > 50$.

The dimensionality $d$ of $y$ and the parameter $K$ for each embedding were determined by estimating the reconstruction quality of the lower-dimensional coordinates. The residual variance $1 - \rho^2_{\mathcal{M},y}$ between the distances in $\mathcal{M}^{(\cdot)}$ and the pairwise distances of $y^{(\cdot)}$ was considered. For each embedding step ($y^e$, $y^f$ and $y^J$), we determined the combination of $K$ and $d$ that minimised the residual variance. Grid-search parameters were set to $d^* \in [1,5]$ and $K^* \in [5,100]$. Final parameters were $K = [50,30,45]$ and $d = [5,5,4]$ for $y^e$, $y^f$ and $y^J$. We considered a model of the disease distributions ($y^e$, $y^f \rightarrow y^{c_1}$) and a model also including the deformation ($y^e$, $y^f$, $y^J \rightarrow y^{c_2}$). Parameters for both models were $K_{c_1} = 55$ and $K_{c_2} = 60$ with $d_{c_1} = 4$ and $d_{c_2} = 4$.

**Table 1.** Pearson correlation coefficient between the first three embedding coordinates and the distributions using the median ($\varphi$), median absolute deviation ($\rho$), skewness ($\gamma_1$), kurtosis ($\gamma_2$). [$* = p < 0.05$, $\dagger = p < 10^{-3}$]

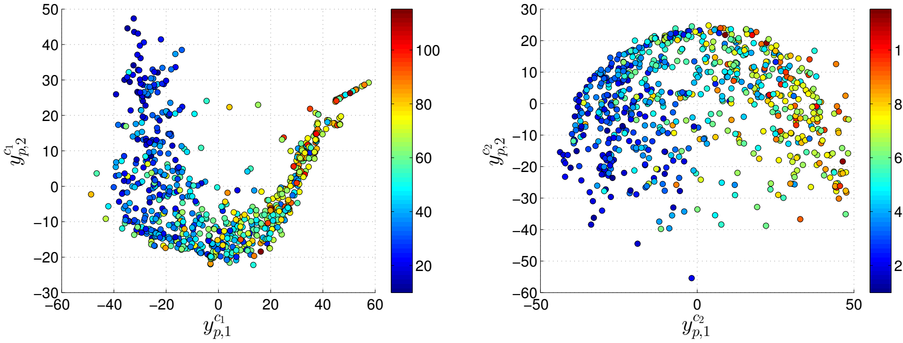
| | PRM$_{emph}$ | | | PRM$_{fSAD}$ | | | $J$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $y^e_1$ | $y^e_2$ | $y^e_3$ | $y^f_1$ | $y^f_2$ | $y^f_3$ | $y^J_1$ | $y^J_2$ | $y^J_3$ |
| $\varphi$ | 0.96$^\dagger$ | -0.19$^\dagger$ | 0.01 | 0.97$^\dagger$ | 0.07 | -0.01 | -0.48$^\dagger$ | -0.06 | 0.04 |
| $\rho$ | 0.89$^\dagger$ | 0.22$^\dagger$ | -0.00 | 0.35$^\dagger$ | -0.36$^\dagger$ | -0.41$^\dagger$ | -0.46$^\dagger$ | 0.14$^*$ | -0.09 |
| $\gamma_1$ | -0.71$^\dagger$ | -0.28$^\dagger$ | 0.00 | -0.86$^\dagger$ | 0.21$^\dagger$ | 0.16$^\dagger$ | -0.68$^\dagger$ | -0.24$^\dagger$ | 0.00 |
| $\gamma_2$ | -0.41$^\dagger$ | -0.26$^\dagger$ | -0.01 | -0.37$^\dagger$ | 0.33$^\dagger$ | 0.26$^\dagger$ | -0.36$^\dagger$ | -0.18$^\dagger$ | -0.01 |

### 3.2 Associations with disease severity

Correlations between the embeddings and distribution moments were computed (Table 1). The first and second components of the embeddings had strong to moderate correlations with the distribution parameters, demonstrating that manifold learning of the distributions modelled the variation in the population.

We considered several models to predict COPD severity using FEV$_1$%predicted and FEV1/FVC (Table 2). We considered three simple models (mean PRM$_{emph}$, mean PRM$_{fSAD}$ and mean Jacobian $\mu(J)$) and compared them to univariate and multivariate models of embedding coordinates ($y$). The univariate models ($y_1^{(e,f)}$) showed moderate improvement over the simple mean models. However, the combined models ($y_1^{c_1}$ and $y_1^{c_2}$) improved model prediction. The multivariate models demonstrated best performance, with model 2 ($y^{c_2} = y^e + y^f + y^J$) performing best, even after adjusting for an increase in variables. It had a Bayesian Information Criterion ($BIC$) of 620 compared to 625 ($y^{c_1}$) and 633, 650 and 648 for PRM$_{emph}$, PRM$_{fSAD}$ and $\mu(J)$ respectively. The increase in explanatory power was also seen when correlating the first component of the combined models ($y_1^{c_{1,2}}$) with FEV$_1$%predicted. The first components of the combined models had

Pearson coefficients of $r = 0.67, p < 0.001$ and $r = 0.70, p < 0.001$ respectively. Coefficients for the mean models were $r = -0.63, p < 0.001$, $r = -0.50, p < 0.001$ and $r = 0.52, p < 0.001$ respectively. We also used manifold fusion to create a joint model between mean values of $\text{PRM}_{emph}$ and $\text{PRM}_{fSAD}$ and a second with $\text{PRM}_{emph}$, $\text{PRM}_{fSAD}$ and $\mu(J)$. Pairwise mean differences were used to create $\mathcal{M}^{(\cdot)}$. Correlation of the first component was $r = 0.60, p < 0.001$ and $r = -0.65, p < 0.001$ respectively. This corroborated the utility of combining embeddings based on the local distributions ($y_1^{c_2} \rightarrow r = 0.70, p < 0.001$).



(a) $y^{c_1}$

(b) $y^{c_2}$

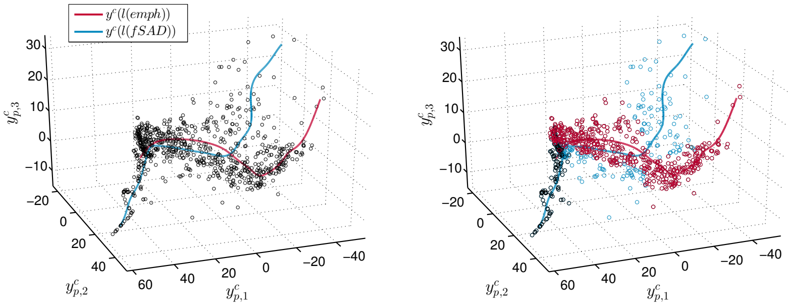
**Fig. 2.** Projection of embeddings a) $y^{c_1}$ and b) $y^{c_2}$ with $\text{FEV}_1$%predicted overlayed.

**Table 2.** Regression of models versus various clinical measures of COPD severity. Model performance quoted as adjusted-$r^2$. [$\dagger = p < 10^{-3}$]

| | Mean features | | | univariate | | | | | multivariate | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | $\text{PRM}_e$ | $\text{PRM}_f$ | $\mu(J)$ | $y_1^{c_1}$ | $y_1^{c_2}$ | $y_1^e$ | $y_1^f$ | $y_1^J$ | $y^{c_1}$ | $y^{c_2}$ | $y^e$ | $y^f$ | $y^J$ |
| $\text{FEV}_1$%p | $0.40^\dagger$ | $0.25^\dagger$ | $0.26^\dagger$ | $0.45^\dagger$ | $\mathbf{0.49}^\dagger$ | $0.42^\dagger$ | $0.29^\dagger$ | $0.13^\dagger$ | $0.48^\dagger$ | $\mathbf{0.51}^\dagger$ | $0.43^\dagger$ | $0.34^\dagger$ | $0.14^\dagger$ |
| $\text{FEV}_1$/FVC | $0.51^\dagger$ | $0.30^\dagger$ | $0.22^\dagger$ | $\mathbf{0.54}^\dagger$ | $0.53^\dagger$ | $0.54^\dagger$ | $0.32^\dagger$ | $0.09^\dagger$ | $0.59^\dagger$ | $\mathbf{0.60}^\dagger$ | $0.55^\dagger$ | $0.38^\dagger$ | $0.10^\dagger$ |

### 3.3 Trajectories of emphysema and fSAD progression

It is likely that trajectories of disease progression in COPD vary depending on the dominant disease phenotype. We assessed whether we can model these in the tissue disease model ($y^{c_1}$). We parameterised $y^{c_1}$ using the emphysema and fSAD distributions as covariates ($l$) with kernel regression: $y^c(l(\cdot)) = \frac{1}{v}\sum_i K(l_i - l)y_i^c$ where $K$ is a Gaussian kernel and $v$ is a normalisation constant. The covariate was the $\mathcal{L}_{EMD}$ between the distributions and an idealised healthy distribution (distribution peak at $v = 0$). The outcome is two trajectories in the manifold space (Fig.3a). The emphysema trajectory can be considered as the path taken when emphysema progression is dominant and vice-versa for fSAD. We classified patients based on these trajectories. A patient is seen to follow an emphysema progression trajectory if it is closest to $y^c(l(emph))$. At the baseline, patients are classified as both emphysema and fSAD subtypes. When considering two sets of patients stratified by trajectory, the explanatory power of the embeddings improved in comparison to $y^{c_1}$ (Table 2). The emphysema regression produced an adjusted-$r^2$ of 0.52 and 0.63 when predicting FEV$_1$%predicted and $^{\text{FEV}_1}/_{\text{FVC}}$ respectively whilst fSAD was 0.45 and 0.62.



(a) $y^{c_1}$

(b) Classified trajectory

**Fig. 3.** a) Three-dimensional projection of $y^{c_1}$ and b) classified trajectories of $y^{c_1}$.

## 4 Discussion and Conclusion

We have presented a method to parameterise distributions of various local features implicated in COPD progression. The disease distributions model local aspects of tissue destruction whilst modelling global properties of heterogeneity and homogeneity. The deformation distribution quantifies the local effect of disease on lung function. Patients exhibiting different mechanisms of tissue destruction can have identical global averages yet can display different disease distributions. These differences are likely to cause differences in local biomechanical properties, which are captured by the deformation distribution.

We have shown that models of the proposed distributions better predict COPD severity than conventional metrics (Table 2). We have shown that embeddings based on distribution dissimilarities have stronger correlations with FEV$_1$%predicted than those learned from mean differences. Both these results suggest that the position of a patient in the manifold space of $y^{c_1}$ or $y^{c_2}$ is critical for assessing COPD. This was observed in the trajectory classification (Fig.3). Determining the trajectory that a patient is following may help inform therapeutic decisions and improve our understanding of COPD progression.

Complexity of the modelling may be increased to model more specific information about lung pathophysiology. Separate manifolds can be produced on a lobar basis. This is likely to further increase the explanatory power of the models since inter-lobar disease metrics correlate with different aspects of physiology. The detection of regional differences in local deformation may add further important information regarding the pathophysiology of a patient.

**Acknowledgements.** This work was supported by the EPSRC under Grant EP/H046410/1 and EP/K502959/1, and the UCLH NIHR RCF Senior Investigator Award under Grant RCF107/DH/2014. It used data (phs000179.v3.p2) from the COPDGene study, supported by NIH Grant U01HL089856 and U01HL089897.

## References

1. Aljabar, P., Wolz, R., Srinivasan, L., Counsell, S.J., Rutherford, M.A., Edwards, A.D., Hajnal, J.V., Rueckert, D.: A combined manifold learning analysis of shape and appearance to characterize neonatal brain development. IEEE transactions on medical imaging 30(12), 2072–86 (2011)
2. Bragman, F., McClelland, J., Modat, M., Ourselin, S., Hurst, J.R., Hawkes, D.J.: Multi-scale Analysis of Imaging Features and Its Use in the Study of COPD Exacerbation Susceptible Phenotypes. In: MICCAI. pp. 417–424 (2014)
3. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1(1), 269–271 (1959)
4. Galbán, C.J., Han, M.K., Boes, J.L., Chughtai, K.A., Charles, R., Johnson, T.D., Galbán, S., Rehemtulla, A., Kazerooni, E.A., Martinez, F.J., Ross, B.D.: CT-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. Nature Medicine 18(11), 1711–1715 (2013)
5. Harmouche, R., Ross, J.C., Diaz, A.A., Washko, G.R., Estepar, R.S.J.: A Robust Emphysema Severity Measure Based on Disease Subtypes. Academic Radiology 23(4), 421–428 (2016)
6. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: MIND: Modality Independent Neighbourhood Descriptor for Multi-Modal Deformable Registration. Medical Image Analysis 16(7), 1423–1435 (2012)
7. Levina, E., Bickel, P.: The earth mover's distance is the Mallows distance: some insights from statistics. Eighth IEEE International Conference on Computer Vision 2, 251–256 (2001)
8. Modat, M., McClelland, J., Ourselin, S.: Lung Registration Using the NiftyReg Package. Medical Image Analysis for the Clinic: A Grand Challenge EMPIRE 10 pp. 33–42 (2010)
9. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine 98(3), 278–84 (2010)
10. Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., Crapo, J.D.: Genetic epidemiology of COPD (COPDGene) study design. COPD 7(1), 32–43 (2010)
11. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval. International Journal of Computer Vision 40(2), 99–121 (2000)
12. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–23 (2000)