Statistical Methodology for Computed Tomography Scans of the Lungs

Sarah M. Ryan, MS

PhD Candidate
Department of Biostatistics and Informatics
Colorado School of Public Health
University of Colorado Anschutz Medical Campus

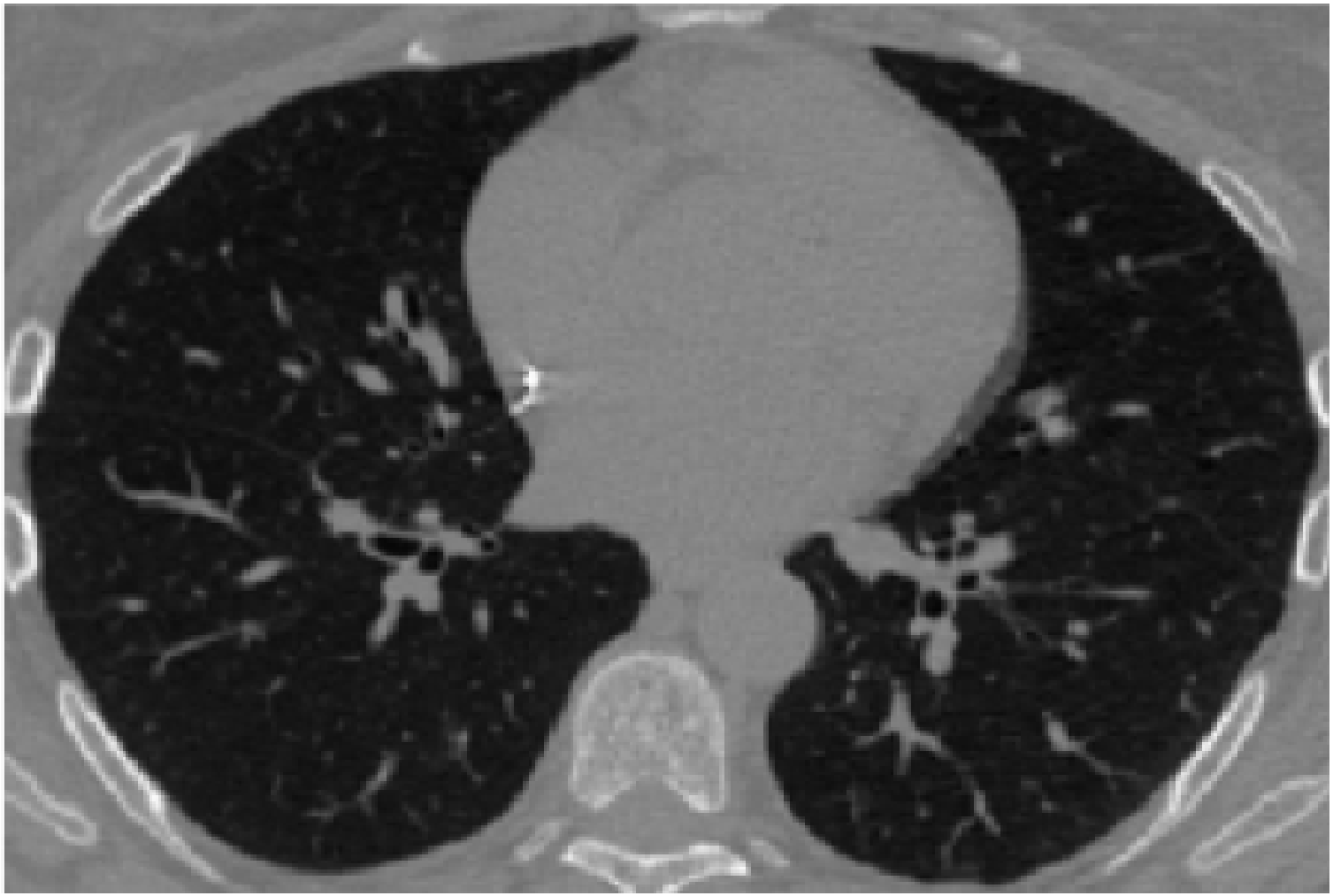
February 25, 2020

Computed Tomography (CT)

A computerized x-ray imaging procedure which generates cross-sectional images of the body that can be combined to form three-dimensional images

CT Scans of the Lung

Healthy Lung

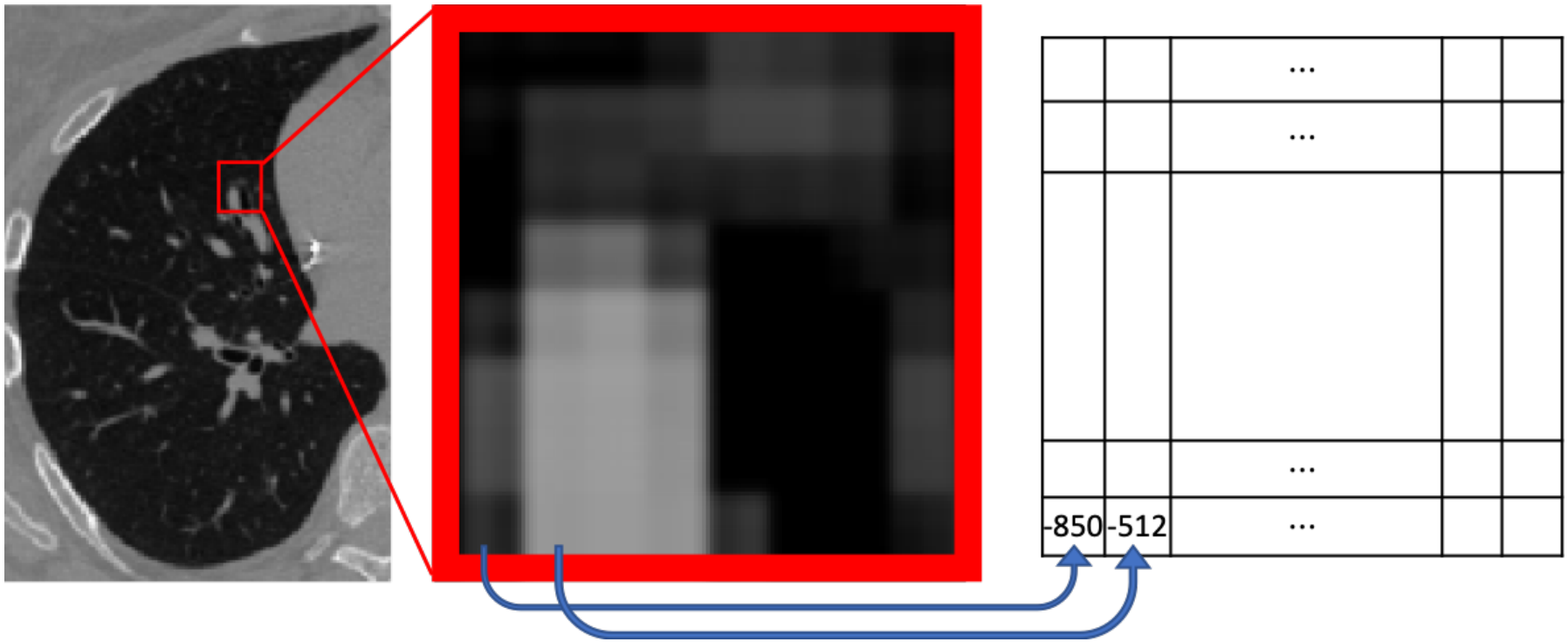


CT Scans of the Lung

Fibrotic Lung

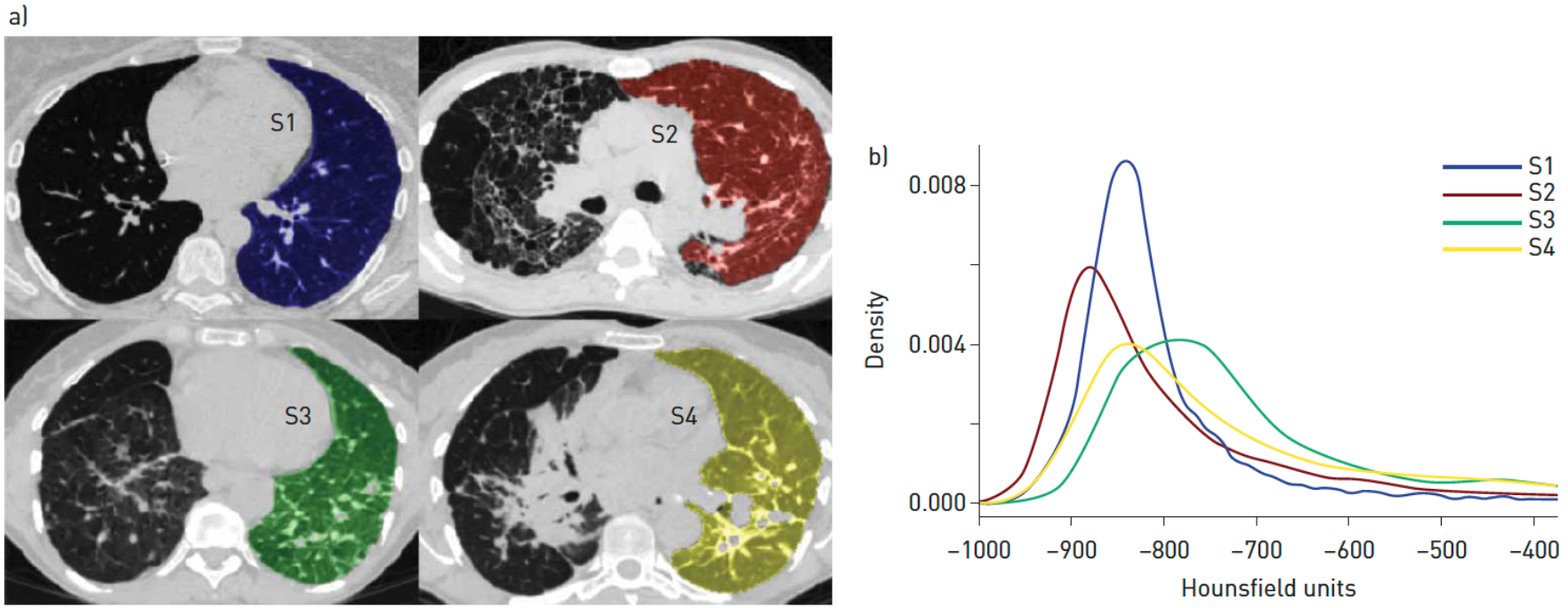
CT Scans are Data

Hounsfield Unit (HU): A measure of the radiodensity of a pixel from a CT scan



Existing Methods for Image Analysis

Radiomics: An emerging field in which large numbers of quantitative features are computed from medical images, providing a rapid, objective, and sensitive quantification of lung abnormalities [Ryan et al., 2019a]



colorado school of public health
UNIVERSITY OF COLORADO
COLORADO STATE UNIVERSITY
UNIVERSITY OF NORTHERN COLORADO

Clinical Question:

Where is disease commonly found in the lung?

Methodological Questions:

1. **How do we align spatial coordinates across scans when scans are different sizes and shapes?**
2. **After aligning spatial coordinates across scans, how do we identify significant areas of disease?**

colorado school of public health
UNIVERSITY OF COLORADO
COLORADO STATE UNIVERSITY
UNIVERSITY OF NORTHERN COLORADO

Clinical Question:

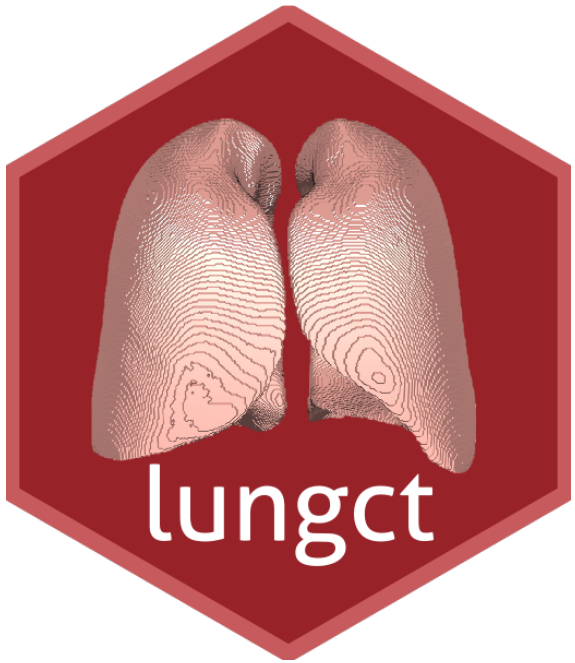
Where is disease commonly found in the lung?

Methodological Questions:

1. How do we align spatial coordinates across scans when scans are different sizes and shapes?
   - **Create a lung template**

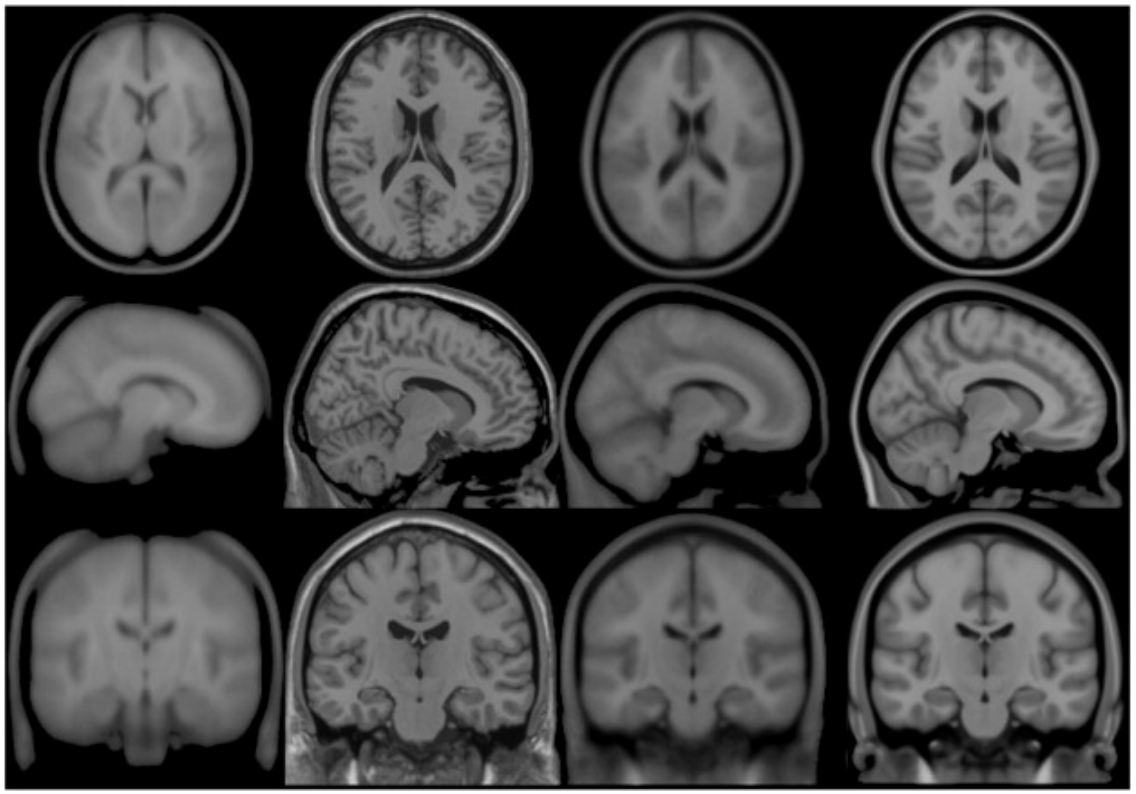
2. After aligning spatial coordinates across scans, how do we identify significant areas of disease?
   - **Develop a spatial model for whole-lung population-level inference**

Create a Lung Template *(with R!)*

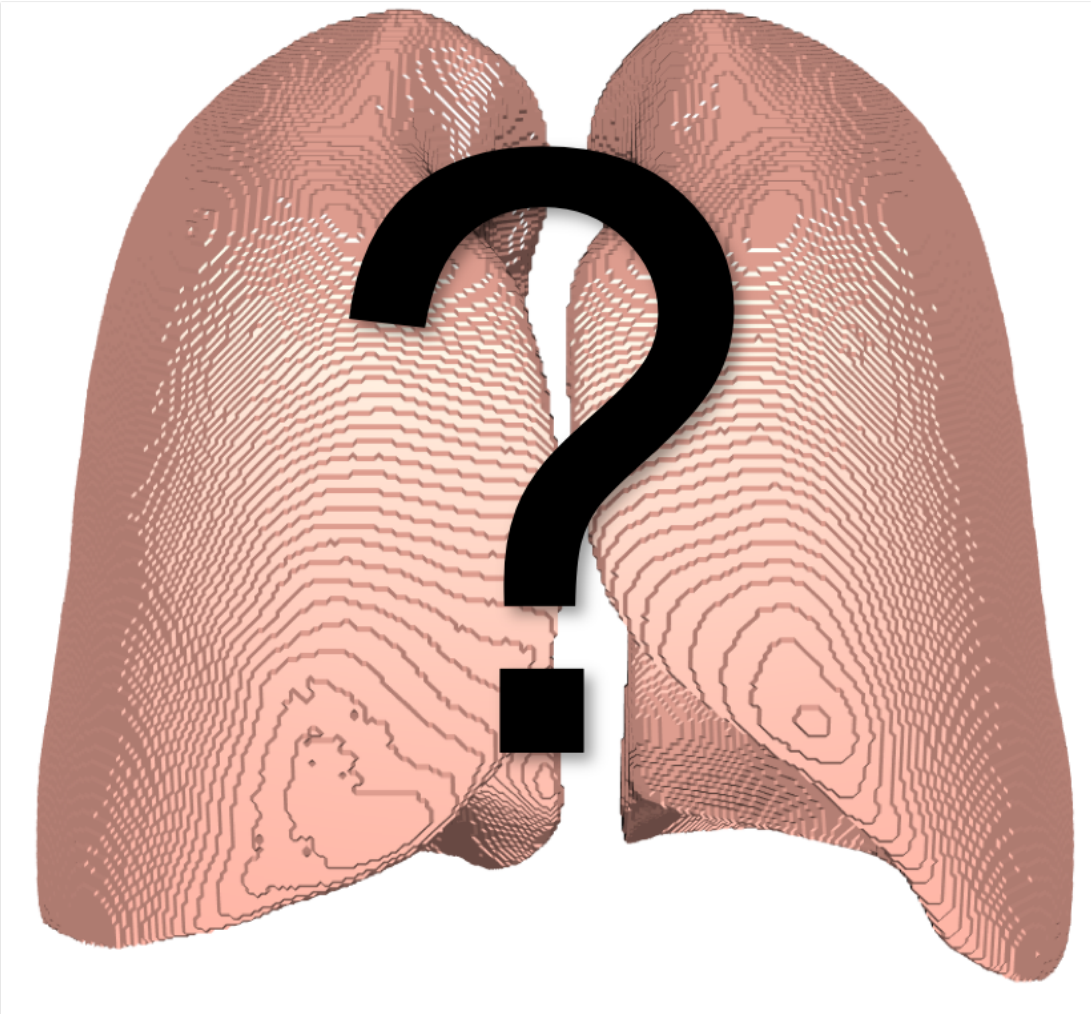


Templates

A **template** is a standardized 3D coordinate frame which allows researchers to combine data across subjects and/or studies [Evans et al., 2012].



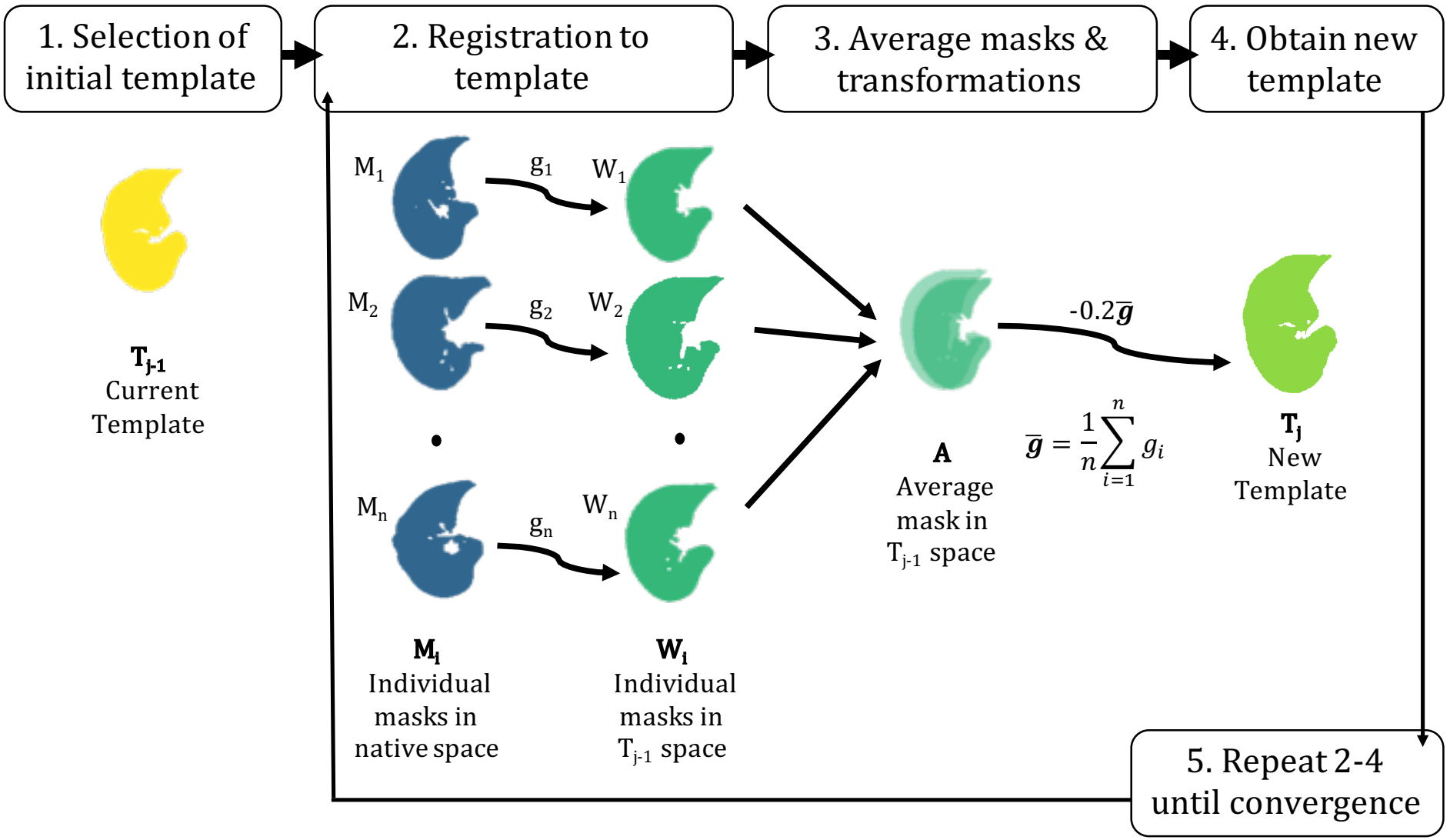
There is no standard lung!



Healthy Population

- **COPDGene** is a cross sectional prospective cohort of smokers with and without COPD, with a goal to perform epidemiological and genetic studies [Regan et al., 2011]
- A small (100) group of similar aged non-smokers without COPD were enrolled to provide a comparison point for CT scans in the aging lung
- We used $\mathbf{N = 62}$ adult subjects, with equal proportions of males/females, age range: 51-72 years.

Template Creation

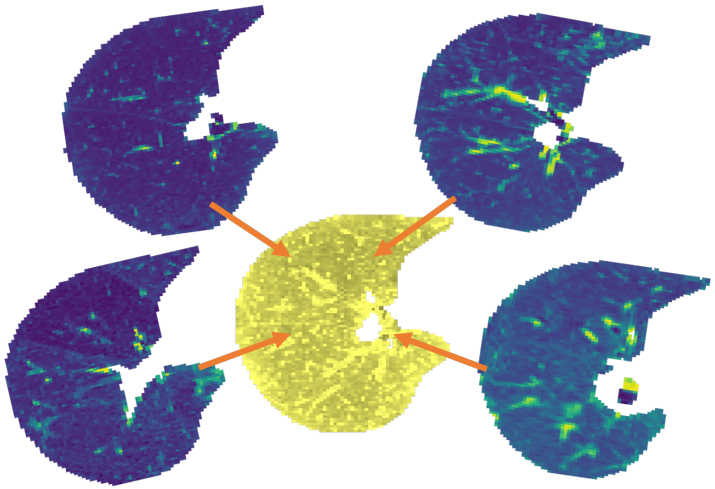


[Ryan et al., 2019b]

Notes on Registration

Registration is the procedure of transforming voxels from their original space into a common space

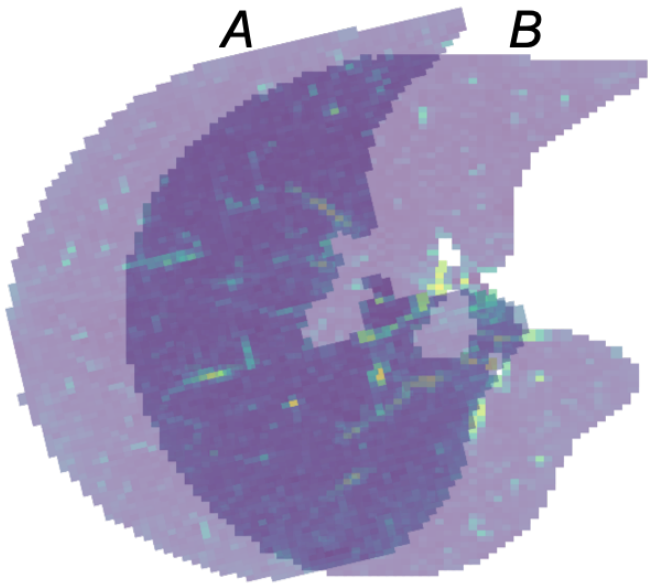
- Left and right lungs were registered separately to account for differences in lung shape and size
- Symmetric Normalization (SyN) non-linear registration was used due to its flexibility and success in EMPIRE10 Challenge [Avants et al., 2008]

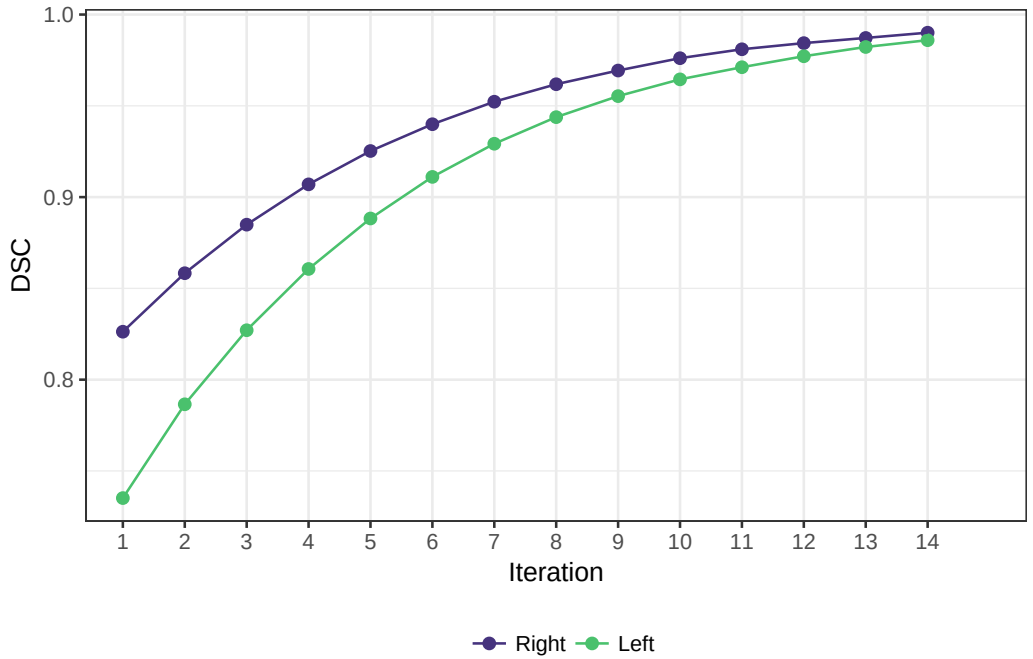


Convergence Criteria

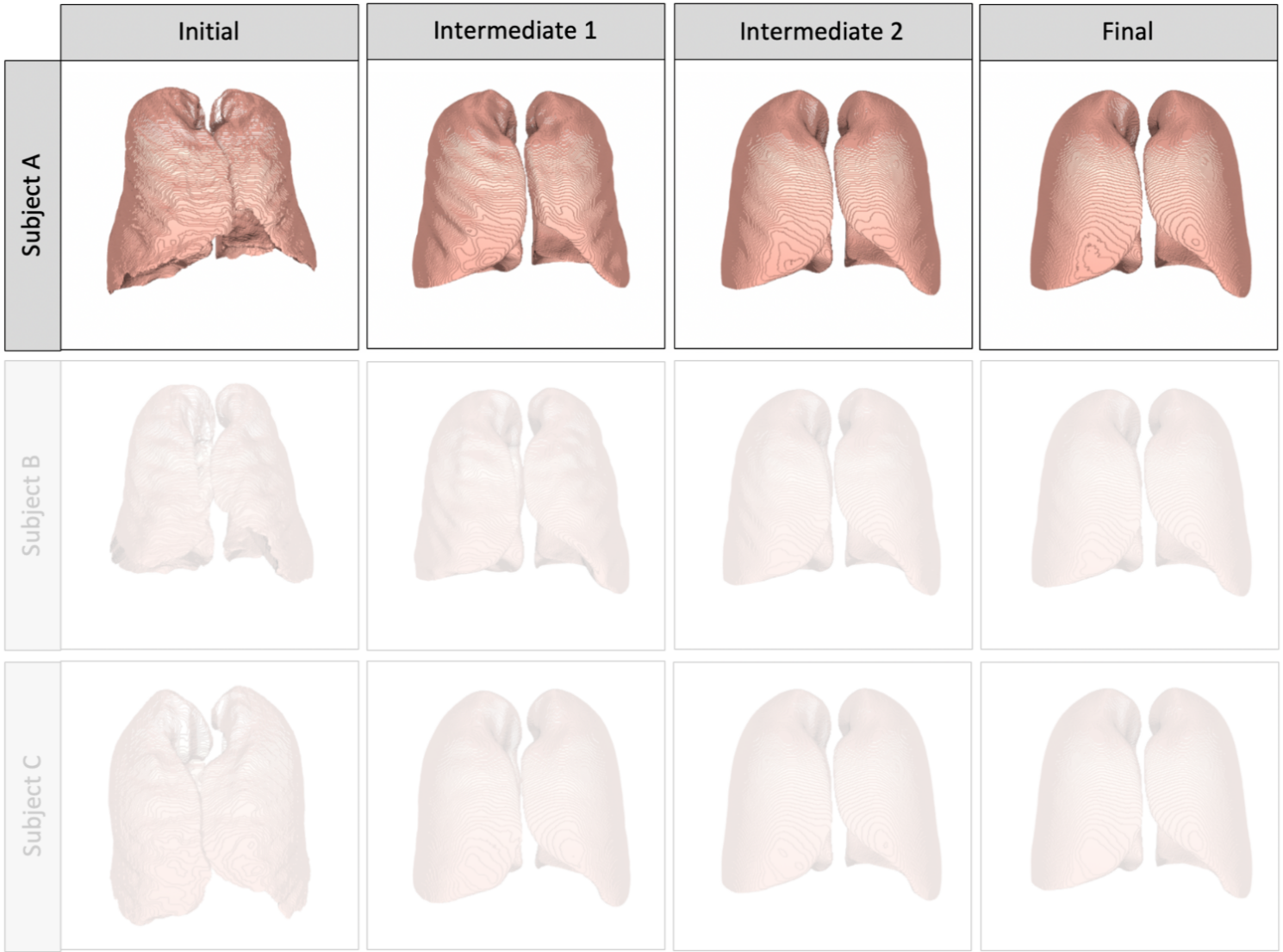
We define convergence using a dice similarity coefficient (DSC) between successive iterations of at least 0.99

$$DSC(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

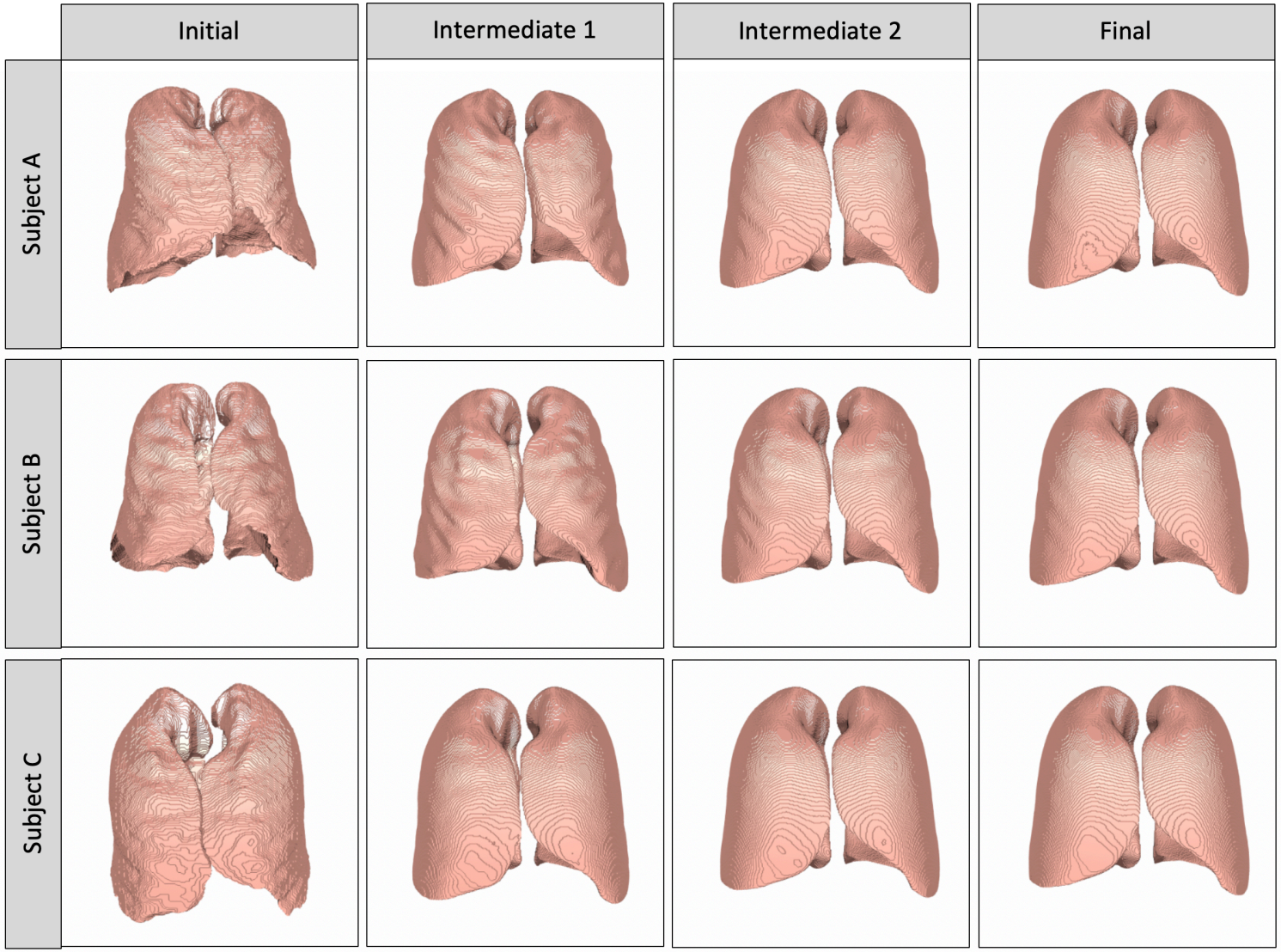




Convergence

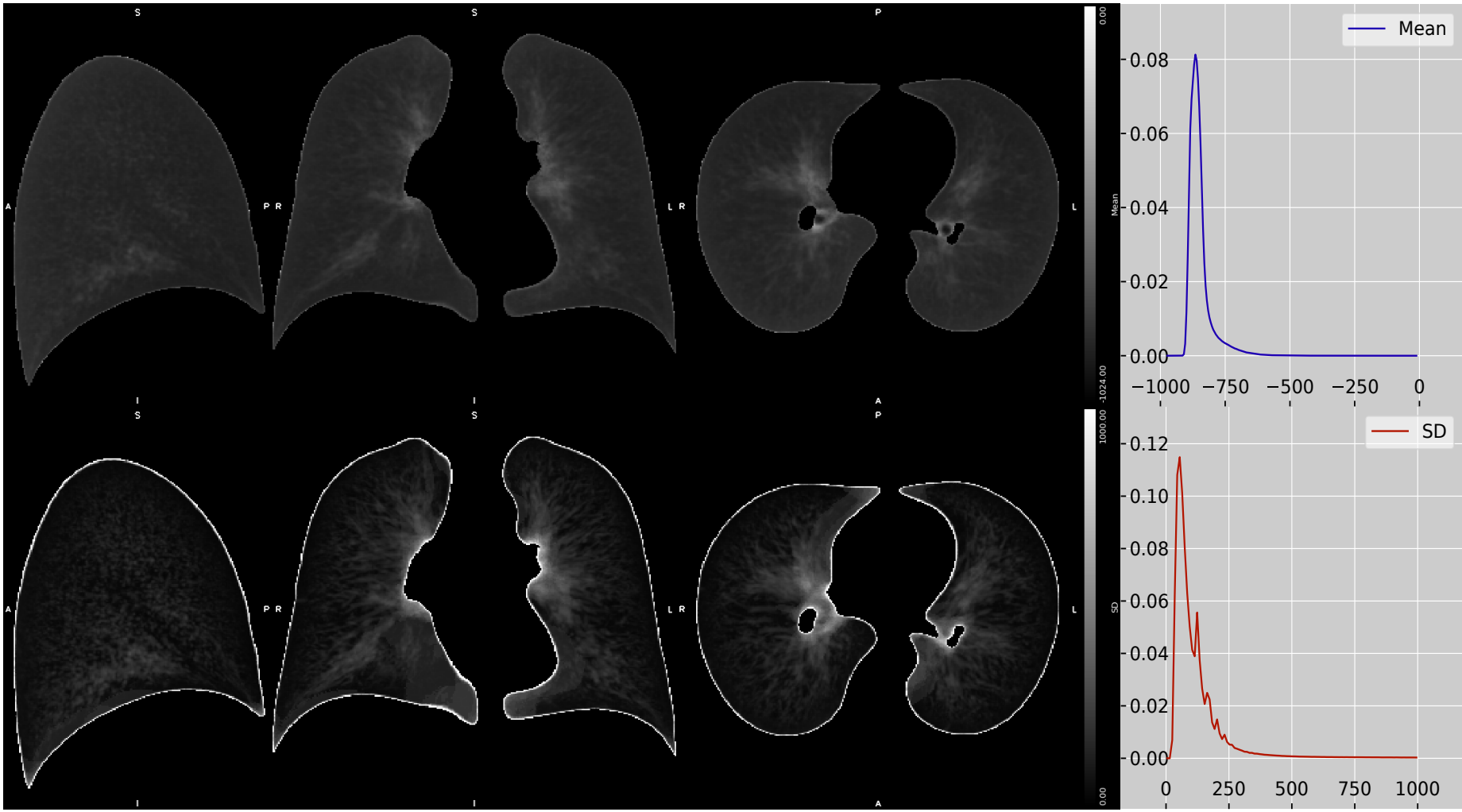


Sensitivity to Initial Template Choice



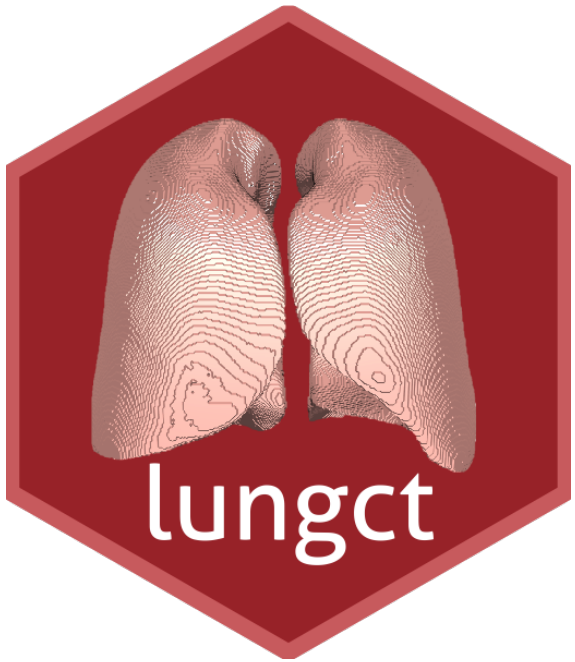
Healthy Lung Template Characteristics

Number of iterations: 14



Conclusions

1. We created the first publicly available standard lung template using healthy adults, which is available for download via *lungct* [Ryan et al., 2019b]
2. We develop a fully-automated and open-source image processing pipeline for lung CTs in R software

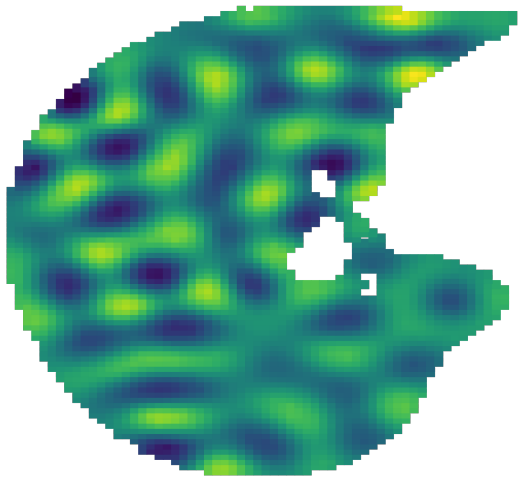


Where is disease commonly found in the lung?

Methodological Questions:

1. How do we align spatial coordinates across scans when scans are different sizes and shapes?
   - ~~**Create a lung template**~~
2. After aligning spatial coordinates across scans, how do we identify significant areas of disease?
   - **Develop a spatial model for whole-lung population-level inference**

Develop a Model for Whole-Lung Population-Level Inference



Classical Method for Imaging Data

Voxel-based morphometry (VBM) - An approach in neuroimaging to obtain differences in intensity values in structural MRI. Traditionally, a mass univariate statistical test is fit on every voxel, known as the general linear model (GLM)

- Fails to account for spatial correlation between voxels
- Arbitrary, post-hoc smoothing of data
- High false positive rate and low power depending on multiple comparison correction

Classical Method for Imaging Data

Voxel-based morphometry (VBM) - An approach in neuroimaging to obtain differences in intensity values in structural MRI. Traditionally, a mass univariate statistical test is fit on every voxel, known as the general linear model (GLM)

- Fails to account for spatial correlation between voxels
- Arbitrary, post-hoc smoothing of data
- High false positive rate and low power depending on multiple comparison correction

A spatial model for high-dimensional data is desired.

Eigenvector Spatial Filtering (ESF)

- A type of low rank approximation
- Describes spatial variation using a linear combination of L basis functions where $L << N$
- Related to Moran's I, a common spatial summary measure
- A computationally-efficient and memory-free approach has been developed [Murakami and Griffith, 2019a]

Eigenvector Spatial Filtering

ESF models spatial maps using eigenvectors associated with the Moran coefficient (MC) [Moran, 1950]:

$$\mathrm{MC}(\mathbf{y}) = \frac{N}{\mathbf{1}'\mathbf{C}\mathbf{1}} \frac{\mathbf{y}'\mathbf{M}\mathbf{C}\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{M}\mathbf{y}} \tag{1}$$

- $\mathbf{y}$ is a spatial response
- $\mathbf{C}$ is a spatial correlation matrix, such as the exponential
- $\mathbf{M}$ is a centering matrix, ensuring eigenvectors are orthogonal and uncorrelated
- $\mathbf{MCM}$ is decomposed into spatially orthogonal variables, or eigenvectors
- A subset of L eigenvectors, $\mathbf{E}$, where $L << N$, are selected

Murakami's ESF Model

The resulting eigenvector matrix, $\mathbf{E}$, can be used to construct spatially-varying coefficients (SVC) in a mixed-effects model [Murakami et al., 2017]:

$$\begin{aligned} \mathbf{y} &= \sum_{k=1}^{K} \mathbf{x}_k \circ \boldsymbol{\beta}_k^{SVC} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}\right) \\ \boldsymbol{\beta}_k^{SVC} &= \beta_k \mathbf{1} + \mathbf{E}\boldsymbol{\gamma}_k, \quad \boldsymbol{\gamma}_k \sim N\left(\mathbf{0}_L, \sigma^2_{k(\gamma)} \boldsymbol{\Lambda}\left(\alpha_k\right)\right) \end{aligned} \tag{2}$$

- Parameters are estimated using restricted maximum likelihood (REML)
- Fits high-dimensional spatial data efficiently
- Shown to have less bias in the SVC estimates as compared to geographically-weighted regression [Murakami and Griffith, 2019a]

Our ESF Model for Imaging Data

Our spatial voxel-based morphometry, or **spVBM**, model incorporates non-spatial random effects $\mathbf{Zb}$. For a single subject i at voxel s, the model is:

$$y_i(s) = \sum_{k=1}^{K} x_{k,i} \circ \beta_k^{SVC}(s) + \mathbf{Z_i b_i} + \varepsilon_i(s) \tag{3}$$

$$\beta_k^{SVC}(s) = \beta_k + \mathbf{E}(s)\gamma_k \tag{4}$$

$$\mathbf{b_i} \sim N(\mathbf{0}, D), \quad \varepsilon_i \sim N\left(0, \sigma_\epsilon^2\right), \quad \gamma_k \sim N\left(\mathbf{0}, \sigma_k^2 \mathbf{\Lambda}(\alpha_k)\right) \tag{5}$$

where $y_i(s)$ is the spatial outcome, $x_{k,i}$ is the k^{th} covariate, $\beta_k^{SVC}(s)$ is a SVC, $\mathbf{Z_i b_i}$ is the random effects term, and $\varepsilon_i(s)$ is the residual component. σ_k^2 controls the spatial variation, α_k is determines the scale of spatial dependence [Murakami et al., 2017].

The spVBM Model: Estimation and Inference

The parameters are estimated in the following steps:

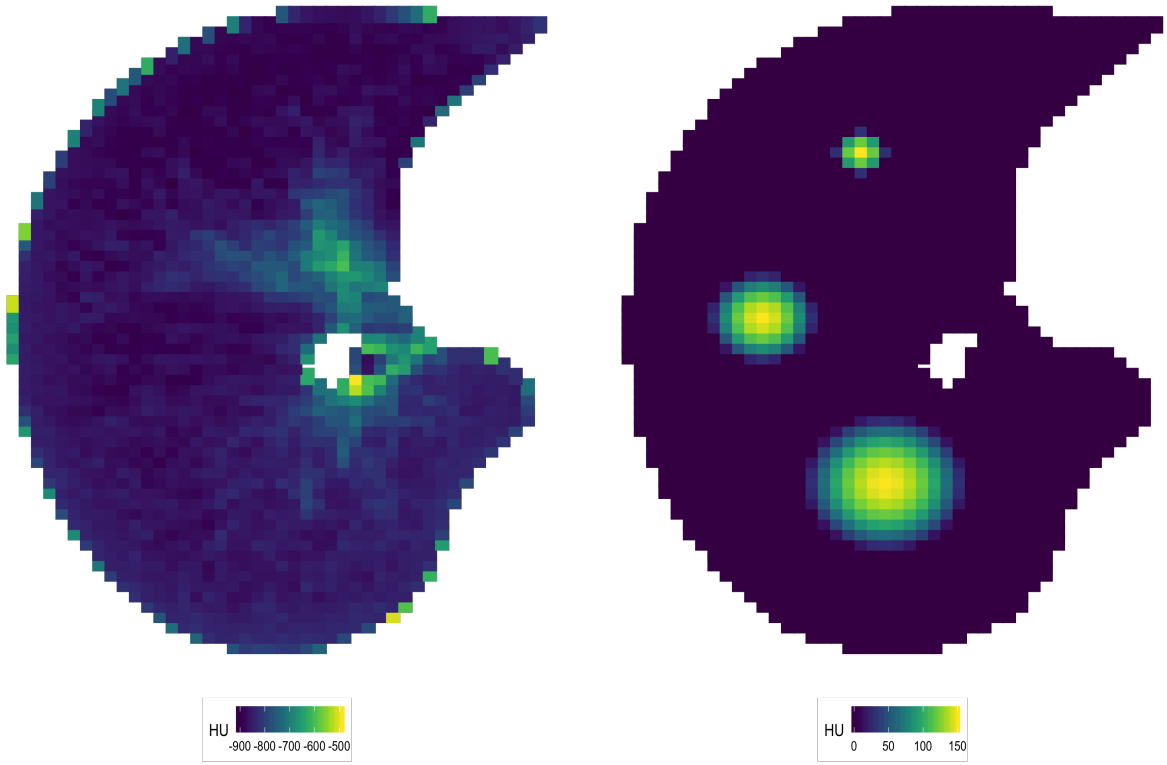
1. β are estimated using maximum likelihood and the normal equations
2. The variance parameters are sequentially estimated by maximizing the restricted log-likelihood

Inference:

- Interested in β_k^{SVC}
- Voxel-level null hypothesis that $\beta_k^{SVC}(s) = 0$
- Wald statistic: $W(s) = \frac{(\beta_k^{SVC}(s)-0)^2}{Var(\beta_k^{SVC}(s))}$

Simulation Study

- Simulate local regions of disease with a binary covariate to represent a case-control study
- Local regions based on small, medium and large disease regions using an Epanechnikov kernel
- Compare spVBM to VBM



Eigenvectors in the Lung

- Distance matrix: (N_u x N_u) dimension
- Eigenvector design matrix: (N_u x L) dimension

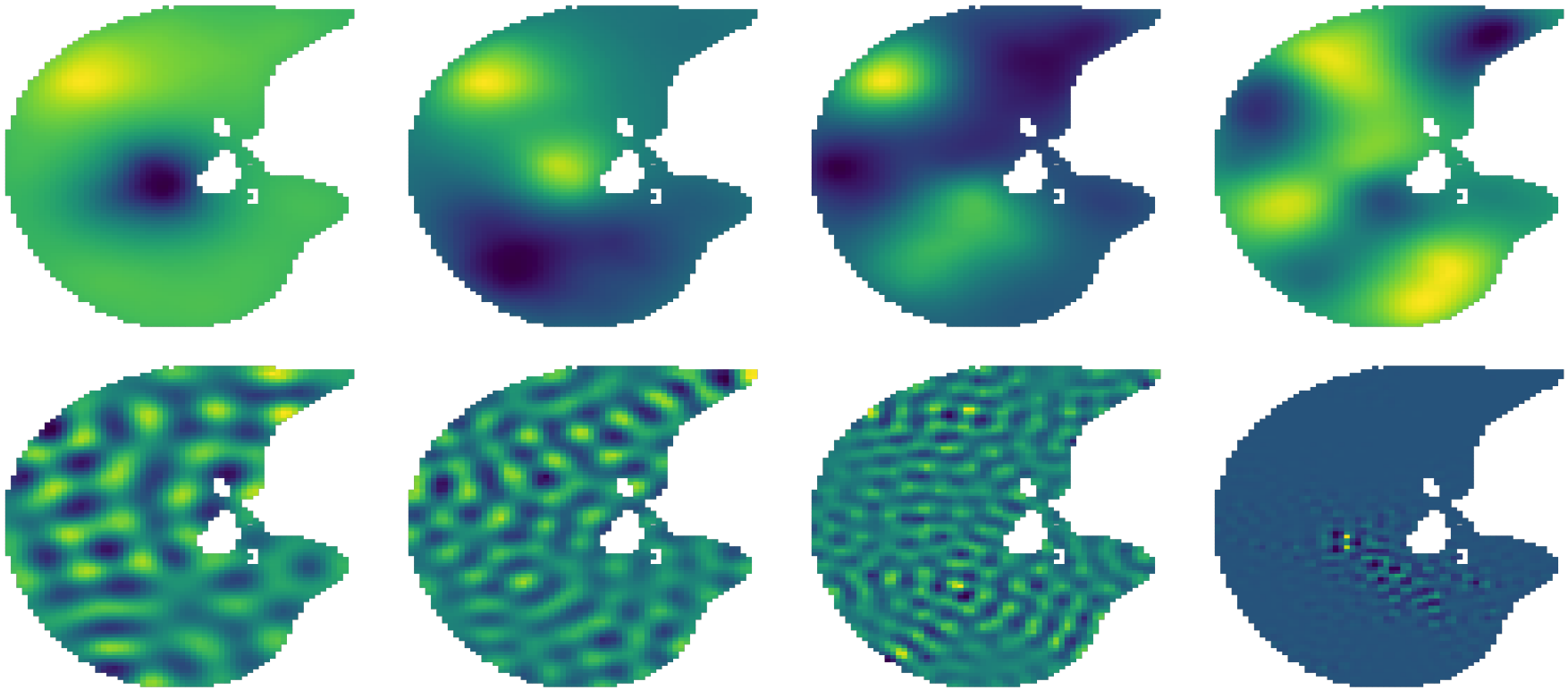
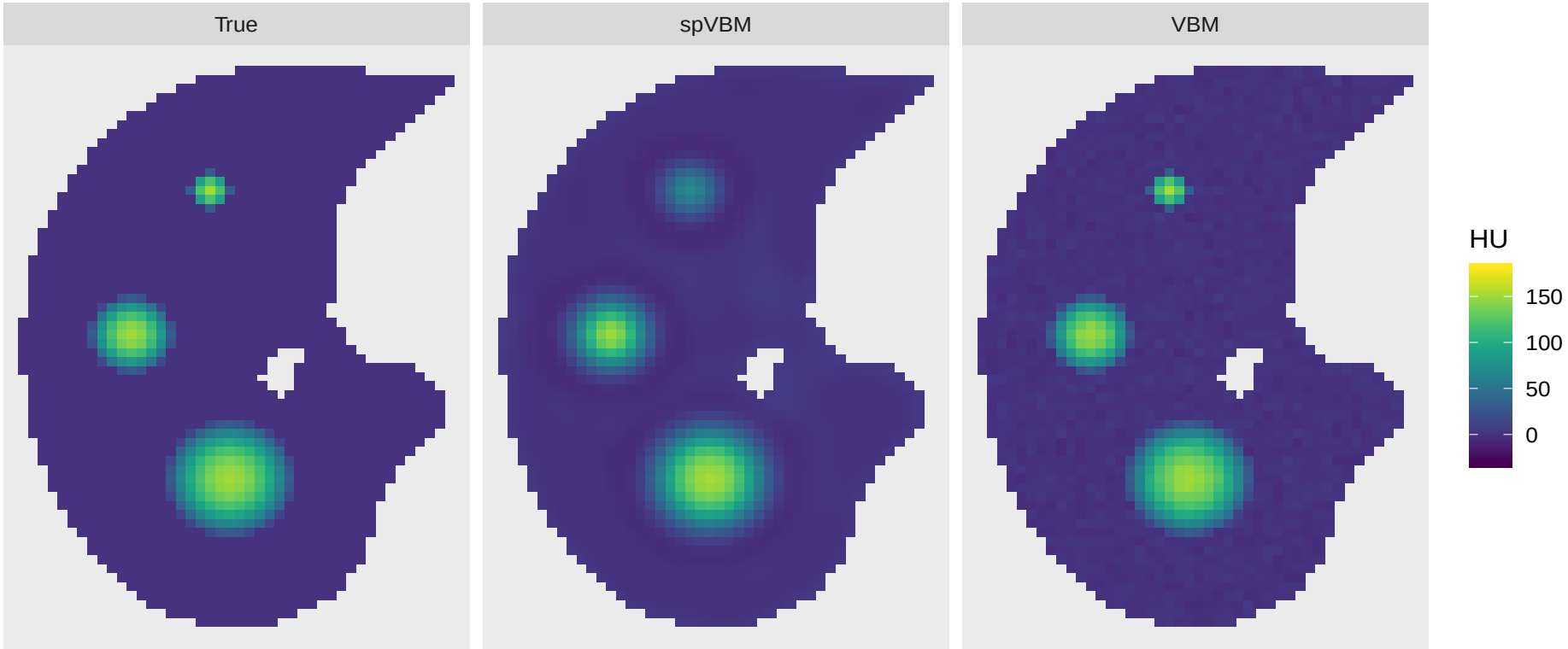
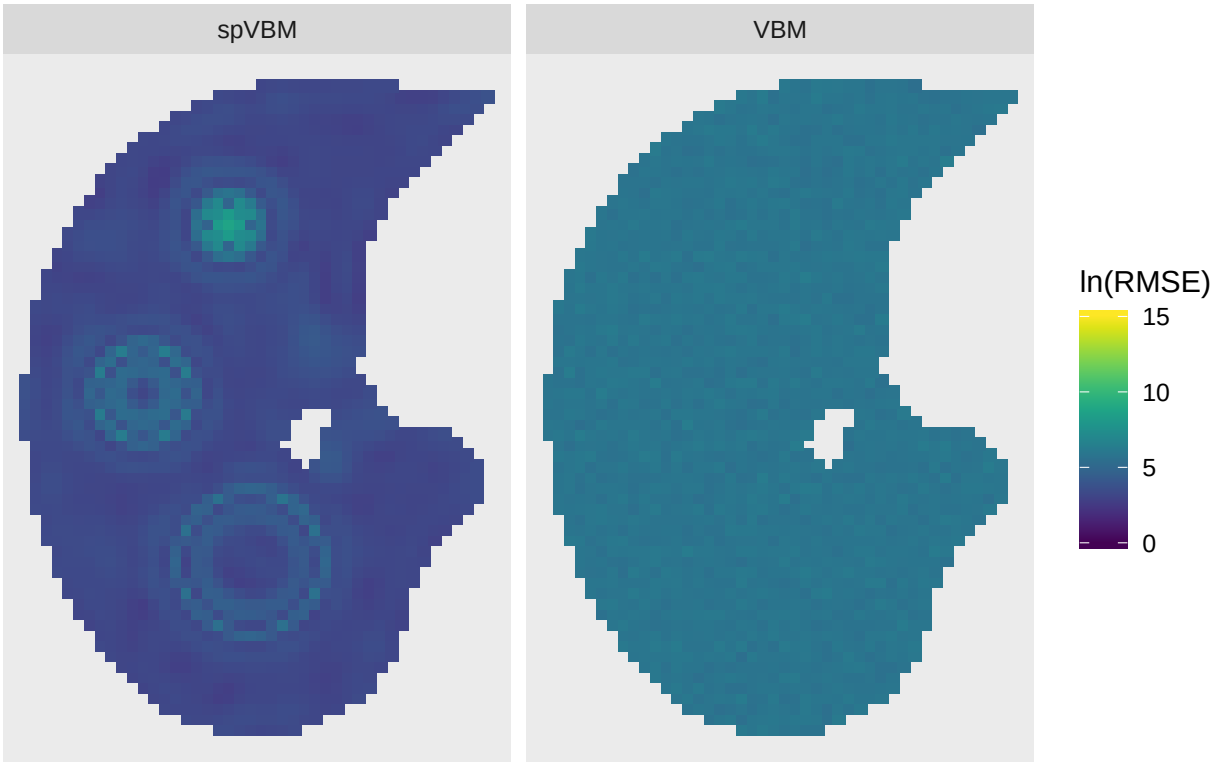


Figure: Moran eigenvectors (1, 2, 4, 8, 100, 200, 400, 800) based on the spatial correlation matrix from a 2D axial slice of the lung.


True
spVBM
VBM
HU
150
100
50
0

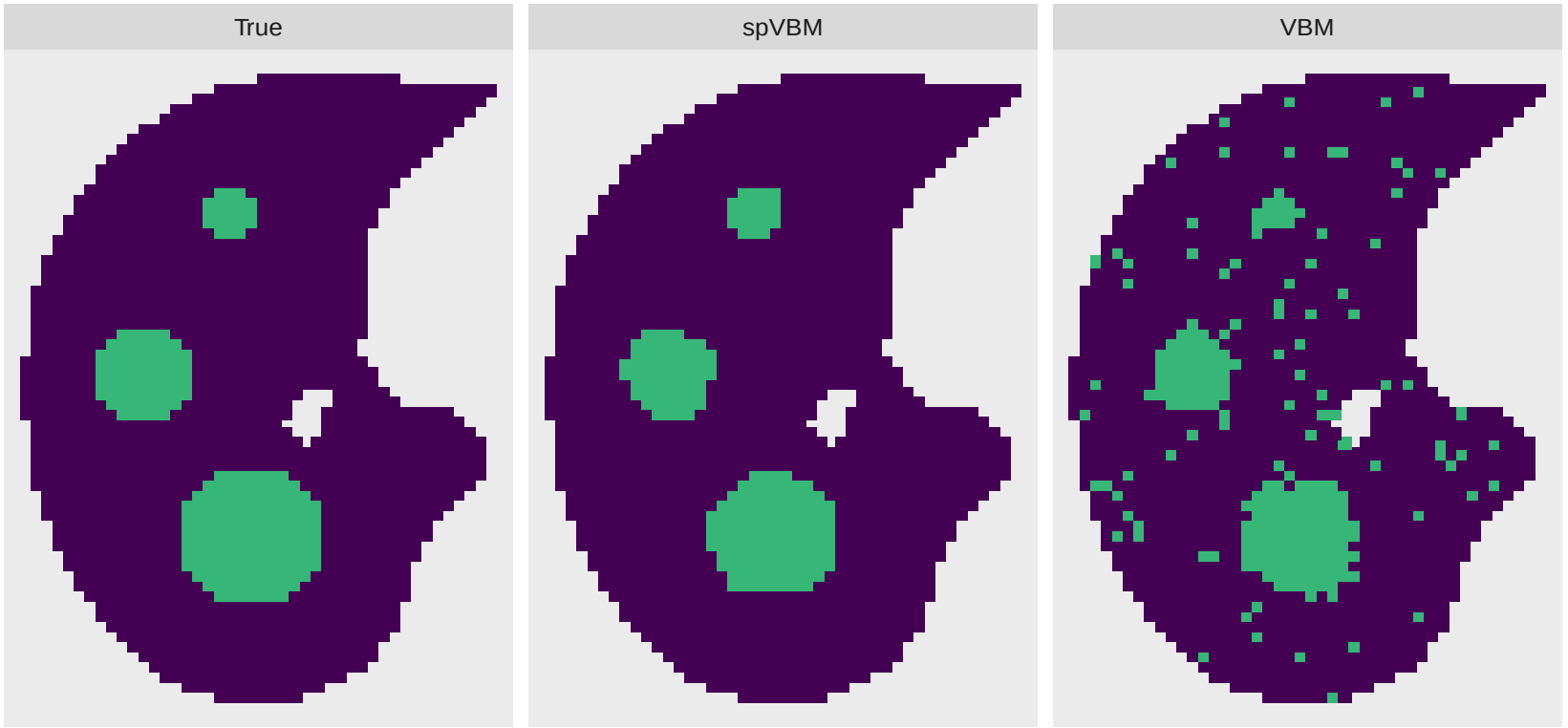
RMSE in $\hat{\beta}_1^{SVC}$



Inference for $\hat{\beta}_1^{SVC}$

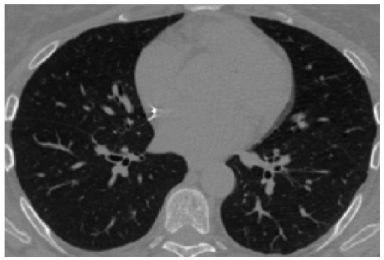
spVBM : FPR = 0.007, FNR = 0.105

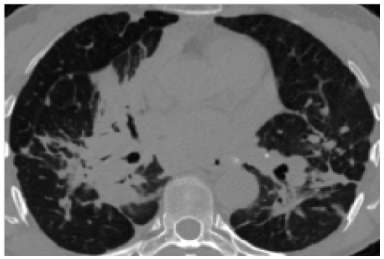
VBM : FPR = 0.050, FNR = 0.315



Application to Diseased Population

GRADS
Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis





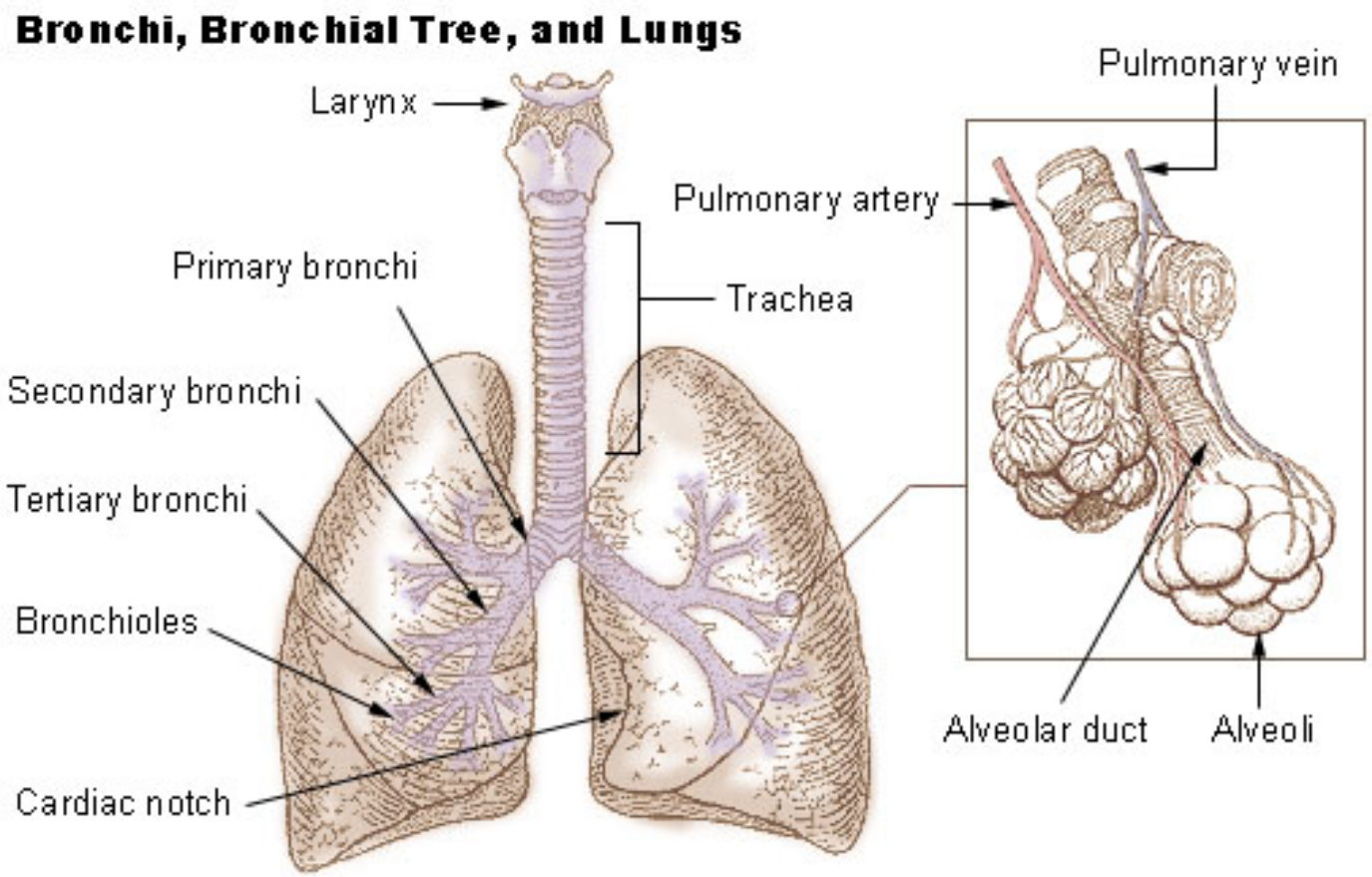
	Healthy	Fibrotic
Sample Size	13	29
Male (%)	6 (46.2)	13 (44.8)
White (%)	11 (84.6)	17 (58.6)
Hispanic (%)	3 (23.1)	1 (3.4)
Age (mean (SD))	56.29 (8.87)	55.03 (7.69)
BMI (mean (SD))	33.09 (4.18)	28.83 (5.38)

Methods Outline

1. Segmented the lungs from the scan
2. Created a study-specific lung template
3. Registered to the template
4. Resampled to $4mm^3$ for computation
   - Right: 29K non-null voxels. Left: 24K non-null voxels
5. Created eigenvector matrix
   - Right: 419 eigenvalues (50 sec). Left: 385 eigenvalues (38 sec)
6. Fit the spVBM model to find the association between fibrosis and HU
   - Right: 538 sec. Left: 384 sec

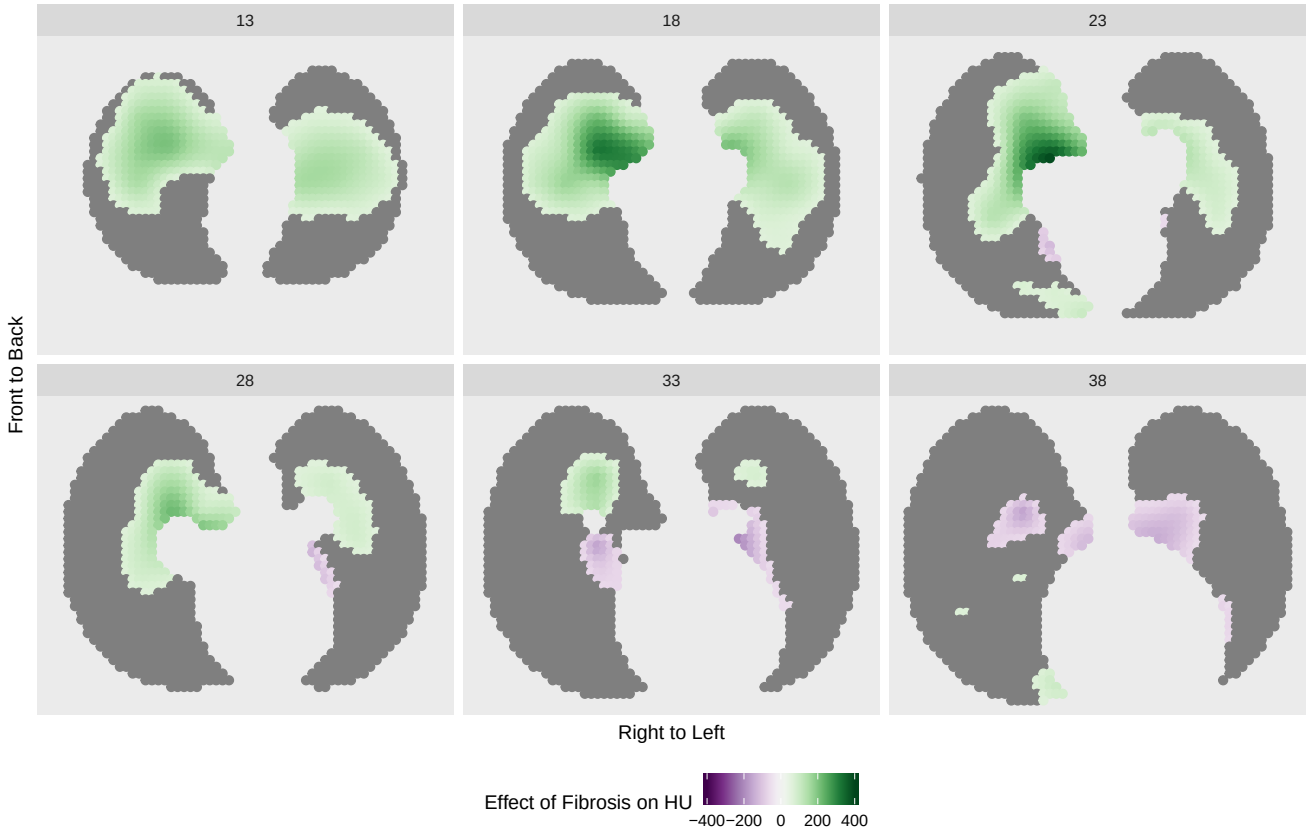
All computations performed on a Mac Pro (Late 2013) 2.7 GHz 12-Core Intel Xeon E5 Processor with 64 GB of memory

Anatomy of the Lung



Results

See in 3D



Compared to healthy patients, patients with fibrosis have areas of significantly higher intensity near the bronchi.

Conclusions

Clinical Question:

Where is disease commonly found in the lung?

Methodological Questions:

1. How do we align spatial coordinates across scans when scans are different sizes and shapes?
   - ~~**Create a lung template**~~
2. After aligning spatial coordinates across scans, how do we identify significant areas of disease?
   - ~~**Develop a spatial model for whole-lung population-level inference**~~

Future Directions

- Develop a spectrum of lung templates
  - Age-specific
  - Disease-specific
  - Compare lung shape and size
- Extend spVBM
  - Compare to other approaches developed for neuroimaging data
  - Apply to structural and functional MRI
  - Estimate parameters in a Bayesian framework
  - Explore the number of eigenvectors
- Deep clustering for imaging data

Collaborators

- Nichole Carlson, PhD, Colorado School of Public Health
- Tasha Fingerlin, PhD, National Jewish Health
- Debashis Ghosh, PhD, Colorado School of Public Health
- Lisa Maier, MD, MSPH, National Jewish Health
- John Muschelli, PhD, Johns Hopkins Bloomberg School of Public Health

Funding Sources

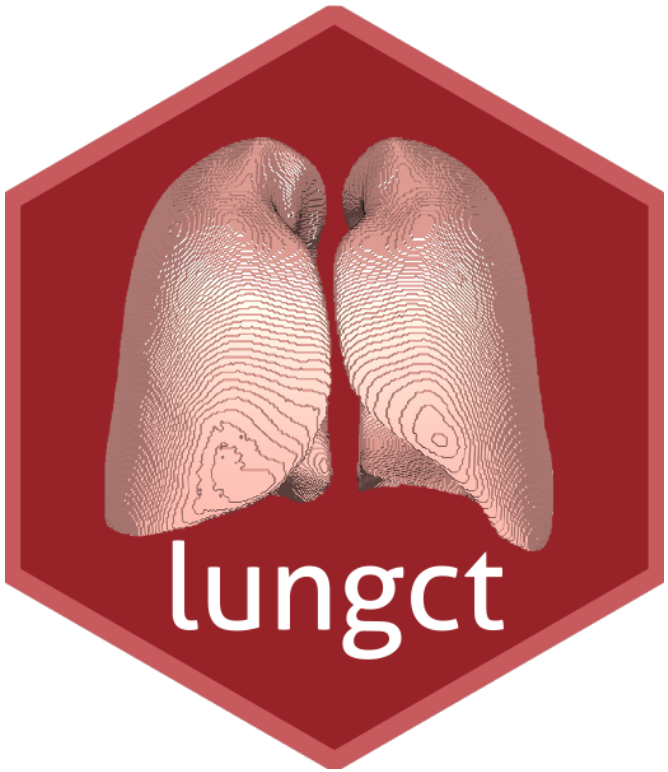
- National Institutes of Health (R01 HL114587; R01 HL142049; U01 HL112695)
- GRADS study (NIH grant U01 HL112707, U01 HL112707, U01 HL112694, U01 HL112695, U01 HL112696, U01 HL112702, U01 HL112708, U01 HL112711, U01 HL112712)
- COPDGene study (NIH grants U01 HL089856 and U01 HL089897, COPD Foundation)

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

Contact

colorado school of public health
UNIVERSITY OF COLORADO
COLORADO STATE UNIVERSITY
UNIVERSITY OF NORTHERN COLORADO

Email: Sarah.M.Ryan@cuanschutz.edu
Twitter: @SarahBiostats
GitHub: ryansar
website: www.SarahMRyan.com



References I

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008).
Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain.
Medical image analysis, 12(1):26–41.

Evans, A. C., Janke, A. L., Collins, D. L., and Baillet, S. (2012).
Brain templates and atlases.
Neuroimage, 62(2):911–922.

Mejia, A. F., Yue, Y., Bolin, D., Lindgren, F., and Lindquist, M. A. (2019).
A bayesian general linear modeling approach to cortical surface fmri data analysis.
Journal of the American Statistical Association, pages 1–26.

Moran, P. A. (1950).
Notes on continuous stochastic phenomena.
Biometrika, 37(1/2):17–23.

Murakami, D. and Griffith, D. A. (2019a).
Eigenvector spatial filtering for large data sets: fixed and random effects approaches.
Geographical Analysis, 51(1):23–49.

Murakami, D. and Griffith, D. A. (2019b).
A memory-free spatial additive mixed modeling for big spatial data.
arXiv preprint arXiv:1907.11369.

Murakami, D., Yoshida, T., Seya, H., Griffith, D. A., and Yamagata, Y. (2017).
A moran coefficient-based mixed effects approach to investigate spatially varying relationships.
Spatial Statistics, 19:68–89.

References II

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., Curran-Everett, D., Silverman, E. K., and Crapo, J. D. (2011).
Genetic epidemiology of copd (copdgene) study design.
COPD: Journal of Chronic Obstructive Pulmonary Disease, 7(1):32–43.

Ryan, S. M., Fingerlin, T. E., Mroz, M., Barkes, B., Hamzeh, N., Maier, L. A., and Carlson, N. E. (2019a).
Radiomic measures from chest hrct associated with lung function in sarcoidosis.
European Respiratory Journal, page 1900371.

Ryan, S. M., Vestal, B., Carlson, N. E., and Muschelli, J. (2019b).
Template creation for high resolution computed tomography scans of the lung in r software.
Academic Radiology.