Histograms should not be used to segment hyperpolarized gas images of the lung

Nicholas J. Tustison, . . . , Jaime F. Mata

Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA

Corresponding author:
Nicholas J. Tustison, DSc
Department of Radiology and Medical Imaging
University of Virginia
ntustison@virginia.edu

Abstract

Magnetic resonance imaging using hyperpolarized gases has facilitated the novel visualization of airspaces, such as the human lung. The advent and refinement of these imaging techniques have furthered research avenues with respect to the growth, development, and pathologies of the pulmonary system. In conjunction with the improvements associated with image acquisition, multiple image analysis strategies have been proposed and developed for the quantification of hyperpolarized gas images with much research effort devoted to semantic segmentation, or voxelwise classification, into clinically-oriented categories based on functional ventilation levels. Given the functional nature of these images and the consequent complexity of the segmentation task, many of these algorithmic approaches reduce the complex spatial image intensity information to intensity-only considerations, particularly those associated with the intensity histogram. Although facilitating computational processing, this simplifying transformation results in the loss of important spatial cues for identifying salient imaging features, such as ventilation defects—an identified correlate of lung pathophysiology. In this work, we demonstrate the interrelatedness of the most common approaches for intensity-only (e.g., histogram), ventilation segmentation of hyperpolarized gas lung imaging for driving voxelwise classification. We evaluate the underlying assumptions associated with each approach and show how these assumptions lead to suboptimal performance. We then illustrate how a convolutional neural network can be constructed in a multi-scale, hierarchically feature-based (i.e., spatial) manner which circumvents the problematic issues associated with existing intensity-only approaches. Importantly, we provide the entire evaluation framework, including this newly reported deep learning functionality, as open-source through the well-known Advanced Normalization Tools (ANTs) library.

1 Introduction

1.1 Early acquisition and development

Early hyperpolarized gas pulmonary imaging research reported findings in qualitative terms.

Descriptions:

- "^{3}He MRI depicts anatomical structures reliably" (1)
- "hypointense areas" (2)
- "signal intensity inhomogeneities" (2)
- "wedge-shaped areas with less signal intensity" (2)
- "patchy or wedge-shaped defects" (3)
- "ventilation defects" (4)
- "defects were pleural-based, frequently wedge-shaped, and varied in size from tiny to segmental" (4)

1.2 Historical overview of quantification

Initial attempts at quantification of ventilation images were limited to ennumerating the number of "ventilation defects" or estimating ventilation defect percentage (as a percentage of total lung volume). Often these measurements were acquired on a slice-by-slice basis.

Prior to the popularization of deep learning in medical image analysis, including in the field of hyperpolarized gas imaging (5), widely used semi-automated or automated segmentation techniques were primarily based on intensity-only considerations. In order of increasing sophistication, these techniques can be categorized as follows:

- binary thresholding based on relative intensities (6, 7),

- linear intensity standardization based on global rescaling of the intensity histogram to a reference distribution based on healthy controls, i.e., "linear binning" (8, 9),
- nonlinear intensity standardization based on piecewise affine transformation of the intensity histogram using the K-means algorithm (10), and
- Gaussian mixture modeling (GMM) with Markov random field (MRF) spatial modeling (11).

The early semi-automated technique used to compare smokers and never-smokers in (6) uses manually drawn regions to determine the mean signal intensity and the standard deviation of the noise to derive a threshold value of three noise standard deviations below the mean intensity. All voxels above that threshold value were considered "ventilated" for the purposes of the study. Related approaches, which continue to be used currently (e.g., (7)), simply use a rescaled threshold value to binarize the segmentation. Similar to the histogram-only algorithms (i.e., linear binning and k-means), these approaches do not take into account the various artefacts associated with MRI such as the non-Gaussianity of the MR imaging noise (12, 13) and the intensity inhomogeneity field (14) which prevent simple intensity thresholds from distinguishing tissue types consistent with that of human experts.

To provide a more granular categorization of ventilation that tracks with clinical qualitative assessment, an increase in the number of voxel classes have been added to the various lung parcellation protocols beyond the binary categories of "ventilated" and "non-ventilated." Linear binning is a simplified intensity standardization approach with six discrete intensity levels (or clusters). The six clusters are evenly spaced throughout the intensity range based on the mean and standard deviation values determined from a cohort of healthy controls.

Intensity rescaling for determination of segmentation clusters of lung images can be thought of as a global affine 1-D transform of the intensity histogram to a standardized 1-D reference histogram. Such a global transform does not account for MR intensity nonlinearities that have been well-studied (15–18) and can cause significant intensity variation even in the same tissue region of the same subject. As stated in (19):

> Intensities of MR images can vary, even in the same protocol and the same

sample and using the same scanner. Indeed, they may depend on the acquisition conditions such as room temperature and hygrometry, calibration adjustment, slice location, B0 intensity, and the receiver gain value. The consequences of intensity variation are greater when different scanners are used.

As we demonstrate in subsequent sections, ignoring these nonlinearities can have significant consequences in the well-studied (and somewhat analogous) area of brain tissue segmentation in T1-weighted MRI (e.g., (20–22)) and we demonstrate its effect in hyperpolarized gas imaging quantification robustness in conunction with noise considerations. In addition, it is not a given that we have a sufficient understanding of what constitutes a "normal" in the context of mean and standard MR intensity values and whether or not those values can be combined in a linear fashion to constitute a reference standard. Of more concrete concern, though, is that the requirement for a healthy cohort for determination of algorithmic parameters introduces an (unnecessary) source of measurement variance.

Previous attempts at histogram standardization (16, 17) in light of these MR intensity nonlinearities have relied on 1-D piecewise affine mappings between corresponding structural features found within the histograms (i.e., peaks and valleys). For example, structural MRI, such as T1-weighted neuroimaging, utilizes the well-known relative intensities of major tissues types (i.e., cerebrospinal fluid, gray matter, and white matter), which characteristically correspond to visible histogram peaks, as landmarks to determine the nonlinear intensity mapping between images. However, in hyperpolarized gas imaging of the lung, no such characteristic structural features exist, generally speaking, between histograms. The approach used by some groups (10) of employing k-means as a clustering strategy (23) to minimize the within-class variance of its intensities can be viewed as an alternative optimization strategy for determining a nonlinear mapping between histograms for a clinically-based MR intensity standardization. Although manual k-means initialization is often used where representative voxels are selected for each class by the operator, linear binning can be considered a type of automated initialization. However, k-means does constitute an algorithmic approach with additional degrees of flexibility over linear binning as it employs basic prior knowledge in the

form of a generic clustering desideratum for optimizing a type of MR intensity standardization[1]

Histogram-based optimization is used in conjunction with spatial considerations in the approach detailed in (11). Based on a well-established iterative approach originally used for NASA satellite image processing and subsequently appropriated for brain tissue segmentation in T1-weighted MRI (24), a GMM is used to model the intensity clusters of the histogram with class modulation in the form of probabilistic voxelwise label considerations within image neighborhoods using the expectation-maximization algorithm. Initialization for this particular application is in the form of k-means clustering which, itself, is initialized automatically using evenly spaced cluster centers—similar to linear binning without the reference distribution. This has a number of advantages in that it accommodates MR intensity nonlinearities, like k-means, but in contrast to k-means and the other algorithms outlined, does not use hard intensity thresholds for distinguishing class labels. However, as we will demonstrate, this algorithm is also flawed in that it implicitly assumes, incorrectly, that meaningful structure is found, and can be adequately characterized, within the associated image histogram in order to optimize class labeling.

Additionally, many of these segmentation algorithms use the N4 bias correction preprocessing algorithm (25) to mitigate MR intensity inhomogeneity artefacts which is an extension of the popular nonparametric nonuniform intensity normalization (N3) algorithm (14). Interestingly, N3/N4 also iteratively optimizes towards a final solution using information from both the histogram and image domains. Based on the intuition that the bias field acts as a smoothing convolution operation on the original image intensity histogram, N3/N4 optimizes a nonlinear intensity mapping, based on histogram deconvolution, which smoothly varies across the image. This nonlinear mapping sharpens the histogram peaks which presumably correspond to tissue types. While such assumptions are appropriate for the domain in which N3/N4 was developed (i.e., T1-weighted brain tissue segmentation) and while it is assumed that the enforcement of low-frequency modulation of the intensity mapping prevents new image features from being generated, it is not clear what effects N4 parameter choices have on the final segmentation

[1]The prior knowledge for histogram mapping is the general machine learning heuristic of clustering samples based on the minimizing within-class distance while simultaneously maximizing the between-class distance. In the case of k-means, this "distance" is the variance as optimizing based on the Euclidean distance is NP-hard.

solution, particularly for those algorithms that are limited to intensity-only considerations.

1.3 Motivation for current study

All these methods can be described in terms of the intensity histogram. Investigating the assumptions outlined above, particularly those associated with the nonlinear intensity mappings due to both the MR acquisition and inhomogeneity mitigation preprocessing, we became concerned by the susceptibility of the histogram structure to such variations and the potential effects on current clinical measures of interest (e.g., ventilation defect percentage) derived from these algorithms. Figure 1 provides a visualization representing some of the structural changes that we observed when simulating these nonlinear mappings. It is important to notice that even relatively small alterations in the image intensities can have significant effects on the histogram even though a visual, clinically-based assessment of the image can be unchanged.

Ultimately, we are not claiming that these algorithms are erroneous per se. Much of the relevant research has been limited to quantifying differences with respect to ventilation versus non-ventilation in various clinical categories and these algorithms have certainly demonstrated the capacity for advancing such research. However, these issues influence quantitation in terms of core scientific measurement principles such as precision (e.g., repeatability) and bias. In addition, as acquistion and analysis methodologies improve, so should the level of sophistication and performance of the measurement tools. In evaluating and assessing these algorithms, it is important to note that human expertise leverages more than relative intensity values to identify salient, clinically relevant features in images. Fortunately, modern algorithmic paradigms, specifically deep learning, have the potential for leveraging spatial information from the images that surpasses the perceptual capabilities of previous approaches and even rivals that of human raters (26). We introduced such an approach in (5) and further expand on that work for comparison with existing approaches in this work. In the spirit of open science, we have made the entire evaluation framework, including our novel contributions, available within our ANTsR and ANTsPy libraries for both R and Python users, respectively.

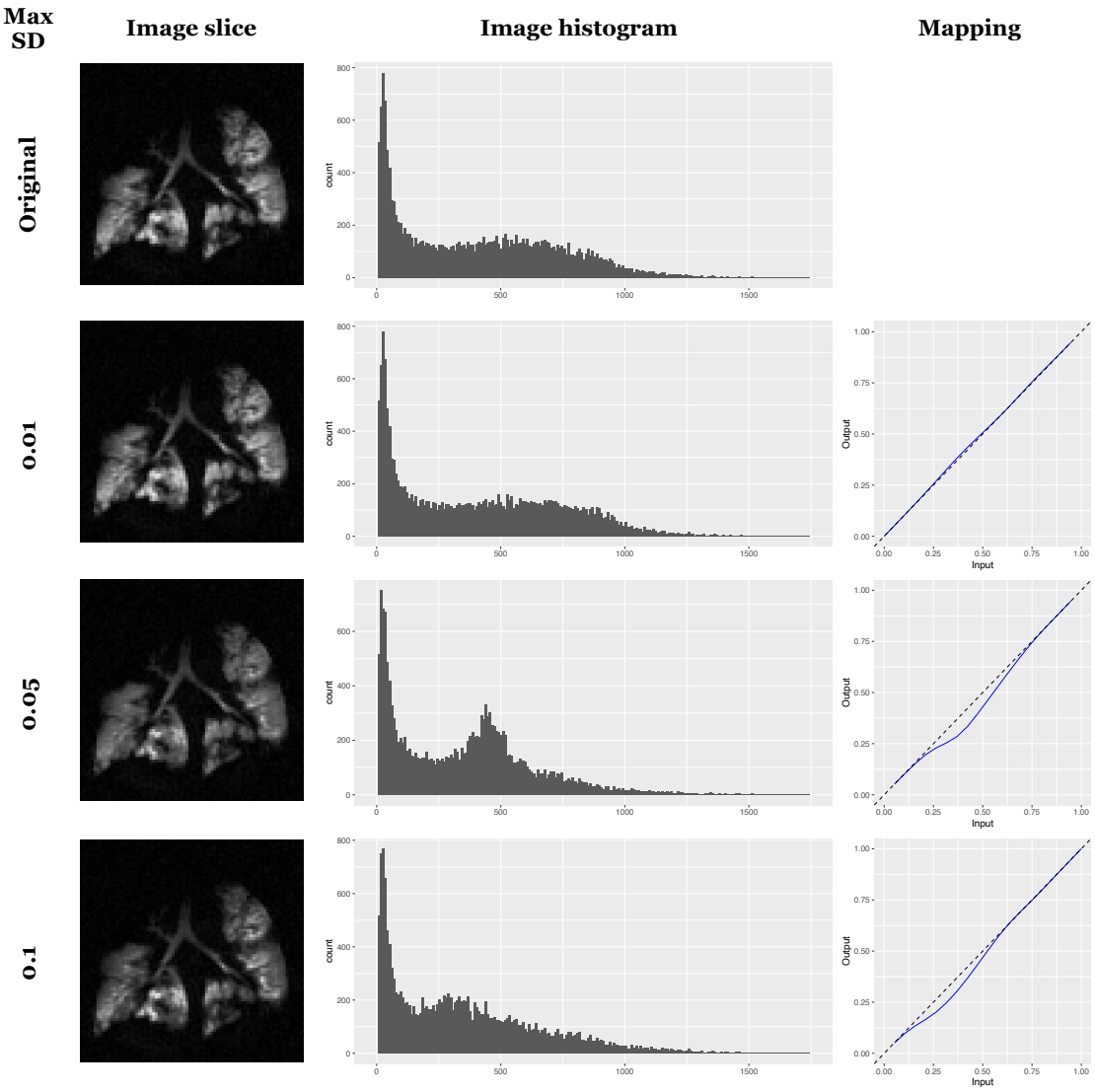

Figure 1: Illustration of the effect of MR nonlinear intensity warping on the histogram structure. We simulate these mappings by perturbing specified points along the bins of the histograms by a Gaussian random variable of 0 mean and specified max standard deviation ("Max SD"). By simulating these types of intensity changes, we can visualize the effects on the underlying intensity histograms and investigate the effects on salient outcome measures. Here we simulate intensity mappings which, although relatively small, can have a significant effect on the histogram structure.

2 Materials and methods

2.1 Image cohort

A retrospective data selection was made

2.2 Algorithmic implementations

All algorithms and evaluation scripts are implemented within the ANTsR/ANTsRNet framework—a component of the ANTsX ecosystem (27) for R users. For the interested reader, ANTsPy/ANTsPyNet make potential evaluation possible with the Python language.

2.3 Introduction of "El Bicho"

We extended the deep learning functionality first described in (5) to improve performance and provide a more clinically granular labeling. In addition, further modifications incorporated additional data during training, added attention gating (28)to the U-net network (29), and novel data augmentation strategies. More details are given below.

2.3.1 Network training

A 2-D per-image-slice U-net model (29) was trained with several parameters recommended by recent U-net exploratory work (30). Four total network layers were employed with 32 filters at the base layer which is doubled at each subsequent layer. Multiple training runs were executed where initial runs employed categorical cross entropy as the loss function. Upon convergence, training continued with a multi-label Dice (31) loss function.

(a) Original. (b) Nonlinear intensity warping. (c) Noise.

Figure 2: Custom data augmentation strategies for training. (b)

Training data (using an 80/20—training/testing split) was composed of the ventilation image along with a lung mask and corresponding ventilation-based parcellation. The ventilation-

based parcellation comprised four labels based on previous experience and the similar choices of other research groups. A total of five random slices per image were selected in the acquisition direction (both axial and coronal) for inclusion within a given batch (batch size = 128 slices). Prior to slice extraction, both random noise and randomly-generated, nonlinear intensity warping was added to the 3-D image (see Figure 2) using the respective ANTsR/ANTsRNet functions:

- `addNoiseToImage`[2] and
- `histogramWarpImageIntensities`[3]

with analogs in ANTsPy/ANTsPyNet. 3-D images were intensity normalized to have 0 mean and unit standard deviation. The noise model was additive Gaussian with 0 mean and a randomly chosen standard deviation value between [0, 0.3]. Histogram-based intensity warping used the default parameters. These data augmentation parameters were chosen to provide realistic but potentially difficult cases for training. In terms of hardware, all training was done on a DGX (GPUs: 4X Tesla V100, system memory: 256 GB LRDIMM DDR4).

2.3.2 Pipeline processing

The proposed deep learning extension was

```
library( ANTsR )
library( ANTsRNet )

# Read in proton and ventilation images.
protonImage <- antsImageRead( "proton.nii.gz" )
ventilationImage <- antsImageRead( "ventilation.nii.gz" )

# Use deep learning lung extraction to get lung mask from proton image.
lungMask <- lungExtraction( protonImage, modality = "proton", verbose = TRUE )

# Run deep learning ventilation-based segmentation
seg <- elBicho( ventilationImage, lungMask, verbose = TRUE )

# Write segmentation and probability images to disk
antsImageWrite( seg$segmentationImage, "segmentation.nii.gz" )
antsImageWrite( seg$probabilityImages[[1]], "probability1.nii.gz" )
antsImageWrite( seg$probabilityImages[[2]], "probability2.nii.gz" )
antsImageWrite( seg$probabilityImages[[3]], "probability3.nii.gz" )
antsImageWrite( seg$probabilityImages[[4]], "probability4.nii.gz" )
```

[2] https://github.com/ANTsX/ANTsR/blob/master/R/addNoiseToImage.R
[3] https://github.com/ANTsX/ANTsRNet/blob/master/R/histogramWarpImageIntensities.R

Listing 1: ANTsR/ANTsRNet command calls for processing a single ventilation image.

2.4 Multi-prong exploratory evaluation

- Brain analogy
  - Show labeled histograms of manually traced images to demonstrate that hard threshold values are inadequate.
- Show how the mean and standard deviation values for linear binning parameters vary based on selection of "normal" cohort.
- Measurement variance based on noise + intensity nonlinearities

Measurement variance based on "normal" cohort selection

Using the ten normals

3 Results

4 Discussion

We recognize that alternative deep learning strategies (hyperparameter choice, training data selection, etc.) could provide comparable and even superior performance to what was presented. However, that is precisely our point—deep learning, generally, presents a much better alternative than what is used currently and we hope that this motivates the field to explore such possible improvements.

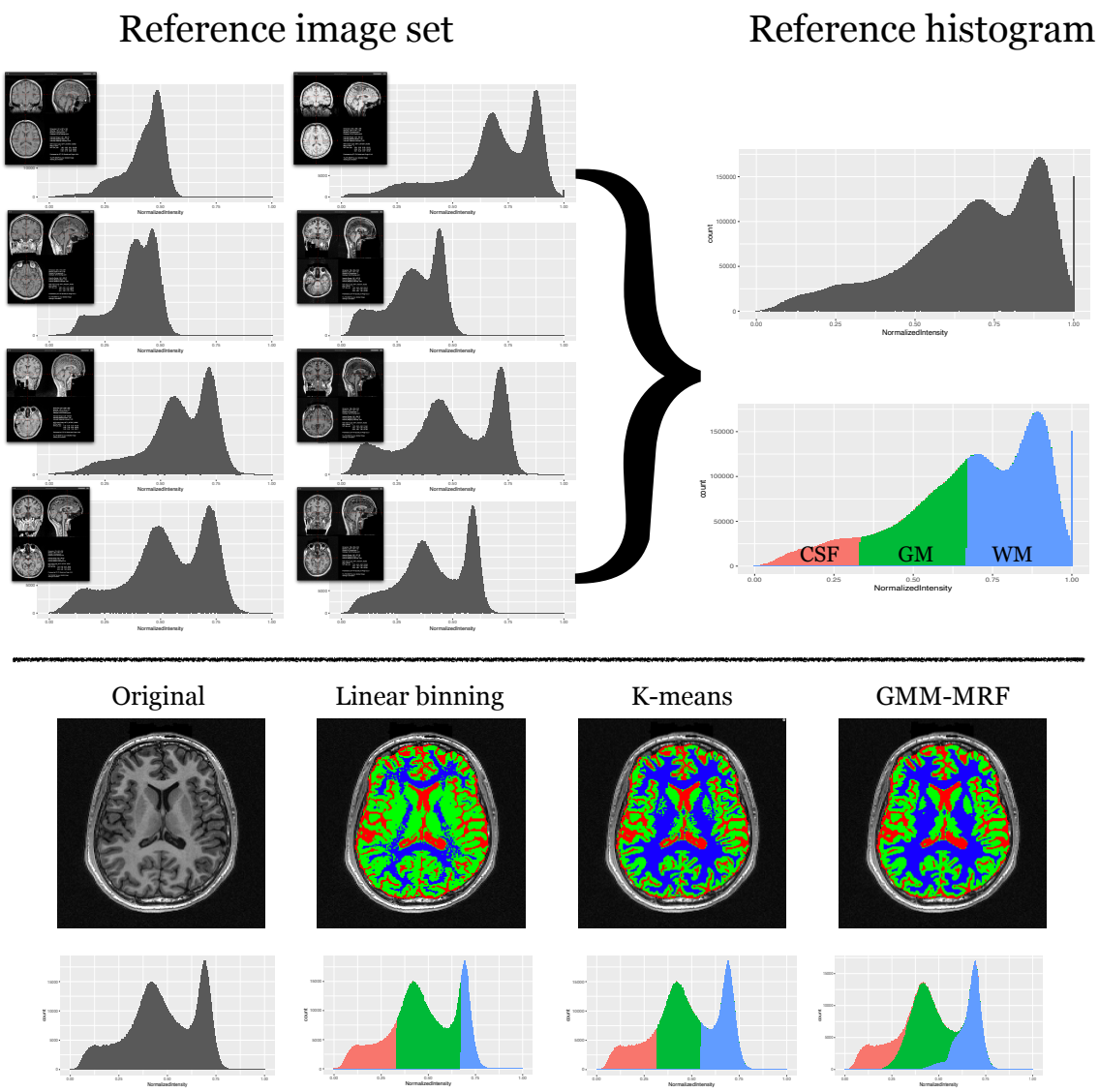

Figure 3

References

1. Bachert P, Schad LR, Bock M, et al.: Nuclear magnetic resonance imaging of airways in humans with use of hyperpolarized 3He. *Magn Reson Med* 1996; 36:192–6.

2. Kauczor HU, Hofmann D, Kreitner KF, et al.: Normal and abnormal pulmonary ventilation: Visualization at hyperpolarized he-3 mr imaging. *Radiology* 1996; 201:564–8.

3. Kauczor HU, Ebert M, Kreitner KF, et al.: Imaging of the lungs using 3He mri: Preliminary

clinical experience in 18 patients with and without lung disease. *J Magn Reson Imaging*; 7:538–43.

4. Altes TA, Powers PL, Knight-Scott J, et al.: Hyperpolarized 3He MR lung ventilation imaging in asthmatics: Preliminary findings. *J Magn Reson Imaging* 2001; 13:378–84.

5. Tustison NJ, Avants BB, Lin Z, et al.: Convolutional neural networks with template-based data augmentation for functional lung image quantification. *Acad Radiol* 2019; 26:412–423.

6. Woodhouse N, Wild JM, Paley MNJ, et al.: Combined helium-3/proton magnetic resonance imaging measurement of ventilated lung volumes in smokers compared to never-smokers. *J Magn Reson Imaging* 2005; 21:365–9.

7. Shammi UA, D'Alessandro MF, Altes T, et al.: Comparison of hyperpolarized 3He and 129Xe mr imaging in cystic fibrosis patients. *Acad Radiol* 2021.

8. He M, Driehuys B, Que LG, Huang Y-CT: Using hyperpolarized 129Xe mri to quantify the pulmonary ventilation distribution. *Acad Radiol* 2016; 23:1521–1531.

9. He M, Wang Z, Rankine L, et al.: Generalized linear binning to compare hyperpolarized 129Xe ventilation maps derived from 3D radial gas exchange versus dedicated multislice gradient echo mri. *Acad Radiol* 2020; 27:e193–e203.

10. Kirby M, Heydarian M, Svenningsen S, et al.: Hyperpolarized 3He magnetic resonance functional imaging semiautomated segmentation. *Acad Radiol* 2012; 19:141–52.

11. Tustison NJ, Avants BB, Flors L, et al.: Ventilation-based segmentation of the lungs using hyperpolarized (3)He MRI. *J Magn Reson Imaging* 2011; 34:831–41.

12. Gudbjartsson H, Patz S: The rician distribution of noisy mri data. *Magn Reson Med* 1995; 34:910–4.

13. Andersen AH: On the rician distribution of noisy mri data. *Magn Reson Med* 1996; 36:331–3.

14. Sled JG, Zijdenbos AP, Evans AC: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging* 1998; 17:87–97.

15. Wendt RE 3rd: Automatic adjustment of contrast and brightness of magnetic resonance images. *J Digit Imaging* 1994; 7:95–7.

16. Nyúl LG, Udupa JK: On standardizing the mr image intensity scale. *Magn Reson Med* 1999; 42:1072–81.

17. Nyúl LG, Udupa JK, Zhang X: New variants of a method of mri scale standardization. *IEEE Trans Med Imaging* 2000; 19:143–50.

18. De Nunzio G, Cataldo R, Carlà A: Robust intensity standardization in brain magnetic resonance images. *J Digit Imaging* 2015; 28:727–37.

19. Collewet G, Strzelecki M, Mariette F: Influence of mri acquisition protocols and image intensity normalization methods on texture classification. *Magn Reson Imaging* 2004; 22:81–91.

20. Zhang Y, Brady M, Smith S: Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 2001; 20:45–57.

21. Ashburner J, Friston KJ: Unified segmentation. *Neuroimage* 2005; 26:839–51.

22. Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC: An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics* 2011; 9:381–400.

23. Hartigan J, Wang M: A k-means clustering algorithm. *Applied Statistics* 1979; 28:100–108.

24. Vannier MW, Butterfield RL, Jordan D, Murphy WA, Levitt RG, Gado M: Multispectral analysis of magnetic resonance images. *Radiology* 1985; 154:221–4.

25. Tustison NJ, Avants BB, Cook PA, et al.: N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging* 2010; 29:1310–20.

26. Zhang R, Isola P, Efros AA, Shechtman E, Wang O: The unreasonable effectiveness of deep features as a perceptual metric. In *2018 ieee/cvf conference on computer vision and pattern recognition*; 2018:586–595.

27. Tustison NJ, Cook PA, Holbrook AJ, et al.: ANTsX: A dynamic ecosystem for quantitative biological and medical imaging. *medRxiv* 2021.

28. Schlemper J, Oktay O, Schaap M, et al.: Attention gated networks: Learning to leverage salient regions in medical images. *Med Image Anal* 2019; 53:197–207.

29. Falk T, Mai D, Bensch R, et al.: U-net: Deep learning for cell counting, detection, and morphometry. *Nat Methods* 2019; 16:67–70.

30. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH: NnU-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2020.

31. Crum WR, Camara O, Hill DLG: Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans Med Imaging* 2006; 25:1451–61.