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 made possible the novel visualization of airspaces, such as the human lung, which has advanced research into the growth, development, and pathologies of the pulmonary system. In conjunction with the innovations associated with image acquisition, multiple image analysis strategies have been proposed and refined for the quantification of hyperpolarized gas images with much research effort devoted to semantic segmentation, or voxelwise classification, into clinically-oriented categories based on ventilation levels. Given the functional nature of these images and the consequent sophistication of the segmentation task, many of these algorithmic approaches reduce the complex spatial image intensity information to intensity-only considerations, which can be contextualized in terms of the intensity histogram. Although facilitating computational processing, this simplifying transformation results in the loss of important spatial cues for identifying salient image features, such as ventilation defects—a well-studied correlate of lung pathophysiology. In this work, we discuss the interrelatedness of the most common approaches for histogram-based segmentation of hyperpolarized gas lung imaging and evaluate the underlying assumptions associated with each approach demonstrating how these assumptions lead to suboptimal performance, particularly in terms of precision. We then illustrate how a convolutional neural network can be trained to leverage multi-scale spatial information which circumvents the problematic issues associated with these approaches. Importantly, we provide the entire processing and evaluation framework, including the newly reported deep learning functionality, as open-source through the well-known Advanced Normalization Tools ecosystem (ANTsX).

Notes to self:

- Calling CNN “el Bicho” until we can come up a different name.
- Jaime to edit Subsections 2.1?
- Possible co-authors: Tally Altes, Kun Qing, John Mugler, Wilson Miller, James Gee, possibly include Mu He after posting to medrxiv.
- Need five more young healthy subjects.

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

Early attempts at quantification of ventilation images were limited to enumerating the number of ventilation defects or estimating the proportion of ventilated lung (4–6). This early work has evolved to current techniques which can be generally categorized in order of increasing algorithmic sophistication as follows:

- binary thresholding based on relative intensities (7, 8),
- linear intensity standardization based on global rescaling of the intensity histogram to a reference distribution based on healthy controls, i.e., "linear binning" (9, 10),
- nonlinear intensity standardization based on piecewise affine transformation of the intensity histogram using the k-means algorithm (11, 12), and

- Gaussian mixture modeling (GMM) of the intensity histogram with Markov random field (MRF) spatial prior modeling (13).

We purposely couch these algorithms within the context of the intensity histogram for facilitating comparison.

An early semi-automated technique used to compare smokers and never-smokers relied on manually drawn regions to determine a threshold value based on the mean signal and noise values (7). Related approaches, which use a simple rescaled threshold value to binarize the ventilation image into ventilated/non-ventilated regions (14), continue to find modern application (8). Similar to the histogram-only algorithms (i.e., linear binning and k-means, discussed below), these approaches do not take into account the various MRI artefacts such as noise (15, 16) and the intensity inhomogeneity field (17) which prevent hard threshold values from distinguishing tissue types precisely consistent with that of human experts. In addition, to provide a more granular categorization of ventilation for greater compatibility with clinical qualitative assessment, many current techniques have increased the number of voxel classes (i.e., clusters) beyond the binary categories of "ventilated" and "non-ventilated."

Linear binning is a simplified type of MR intensity standardization (18) in which a set of healthy controls, all intensity normalized to [0, 1], is used to calculate the cluster threshold values, based on a single Gaussian model. A subject image to be segmented is then rescaled to this reference histogram (i.e., a global affine 1-D transform). This mapping aligns the cluster boundaries such that corresponding labels have the same clinical interpretation. In addition to the previously mentioned issues associated with hard threshold values, such a global transform does not account for MR intensity nonlinearities that have been well-studied (18–22) and are known to cause significant intensity variation even in the same region of the same subject. As stated in (21):

> 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., (23–25)) and we demonstrate its effect in hyperpolarized gas imaging quantification robustness in conjunction 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 deviation 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 a non-negligible source of measurement variance, as we will also demonstrate.

Previous attempts at histogram standardization (18, 20) 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 (11) of employing k-means as a clustering strategy (26) 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. K-means does constitute an algorithmic approach with additional degrees of flexibility and sophistication 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] Although manual k-means initialization is sometimes used where representative voxels are

[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 intensity variance as optimizing based on the Euclidean distance is NP-hard.

selected for each class by the operator, the linear binning strategy of equally spaced cluster centers is often used as a type of k-means automated initialization.

Histogram-based optimization is used in conjunction with spatial considerations in the approach detailed in (13). 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 (27), a GMM is used to model the intensity clusters of the histogram with class modulation in the form of probabilistic voxelwise label considerations, i.e. Markov Random Field (MRF) modeling, within image neighborhoods (28) using the expectation-maximization (EM) algorithm (29). 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. 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 a multi-class labeling.

Additionally, many of these segmentation algorithms use N4 bias correction (30), an extension of the nonuniform intensity normalization (N3) algorithm (17), to mitigate MR intensity inhomogeneity artefacts. 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 (i.e., deformable) intensity mapping, based on histogram deconvolution. This nonlinear mapping is constrained such that its effects smoothly vary across the image. Additionally, due to the deconvolution operation, 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

solution, particularly for those algorithms that are limited to intensity-only considerations and not robust to the aforementioned MR intensity nonlinearities.

1.3 Motivation for current study

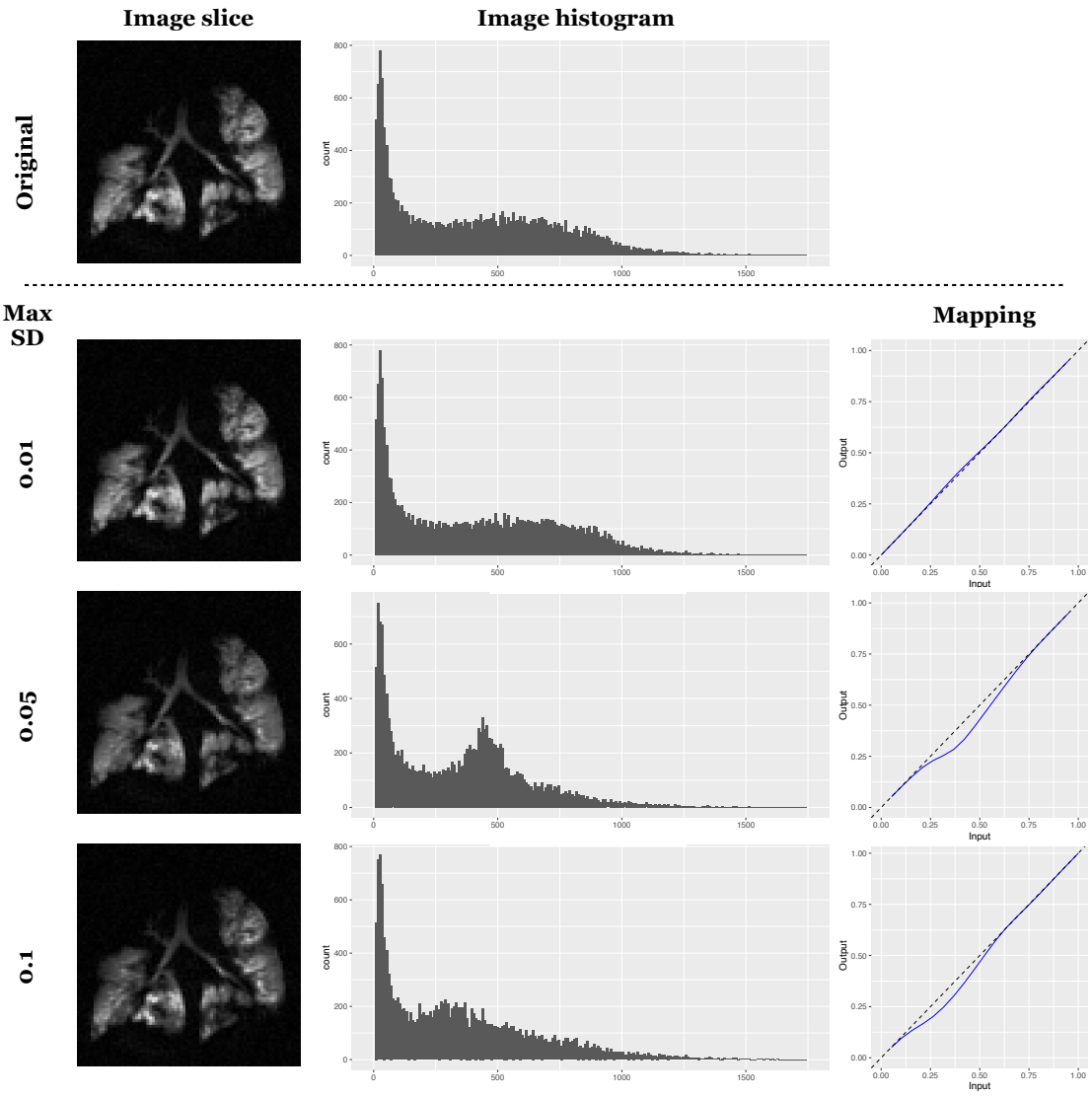


Figure 1: Illustration of the effect of MR nonlinear intensity warping on the histogram structure. We simulate these mappings by perturbing specified points along the bins of the histograms by a Gaussian random variable of 0 mean and specified max standard deviation ("Max SD"). By simulating these types 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.

It should be clear that 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 derived from these algorithms (e.g., ventilation defect percentage). Figure 1 provides a sample 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 remain largely 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., reproducibility and repeatability (31)) and bias which will become more acute as multi-site and large-scale studies are performed. In addition, generally speaking, refinements in measuring capabilities correlates with scientific advancement so as acquisition and analysis methodologies improve, so should the level of sophistication and performance of the measurement tools.

The recent emergence of deep-layered neural networks (32), particularly convolutional neural networks (CNN), is due to their outstanding performance in certain computational tasks, including classification and semantic segmentation in medical imaging (33). Their potential for leveraging spatial information from images surpasses the perceptual capabilities of previous approaches and even rivals that of human raters (34). In assessing these segmentation algorithms for hyperpolarized gas imaging, it is important to note that human expertise leverages more than relative intensity values to identify salient, clinically relevant features in images—something more akin to the complex neural network structure versus the 1-D intensity histogram. We introduced a deep learning approach in (35) 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 Advanced Normalization Tools software ecosystem (ANTsX).

2 Materials and methods

To support the discussion in the Introduction, we perform various experiments to showcase the effects of both MR nonlinear intensity mapping and noise on measurement bias and precision using the popular algorithms described previously, specifically:

- linear binning,
- k-means,
- GMM-MRF, and
- CNN.

We first demonstrate the effects of MR intensity nonlinearities on the analogical application of T1-weighted brain MR segmentation. This evaluation is strictly qualitative as the visual evidence and previous developmental history is overwhelmingly indicative of the need for adequate algorithmic optimization capabilities. We use these qualitative results as a segue to quantifying the effects of the choice of reference cohort on the clustering parameters for the three histogram-based algorithms. We then incorporate the CNN model in exploring additional aspects of measurement variance based on simulating both MR noise and intensity nonlinearities. Finally, we investigate algorithmic accuracy (i.e., bias) in the absence of ground-truth segmentations, by using a clinical diagnostic prediction approach and a study for simultaneous truth and performance level estimation (STAPLE) (36).

A fair and accurate comparison between algorithms necessitates several considerations which have been outlined previously (37). In designing the evaluation study:

- All algorithms and evaluation scripts have been implemented using open-source tools by the first author who is also responsible for the GMM-MRF ("Atropos" in ANTs) and N4 algorithms. The linear binning and k-means algorithms were easily recreated using existing R tools. Similarly, N4, GMM-MRF, and the trained CNN approach are all available through ANTsR/ANTsRNet—`ANTsR::n4BiasFieldCorrection`,

`ANTsR::functionalLungSegmentation`, and `ANTsR::elBicho`.[2]

2.1 Image cohorts

A retrospective dataset was collected consisting of young healthy ($n = 5$), older healthy ($n = 7$), cystic fibrosis (CF) ($n =?$), idiopathic lung disease (ILD) ($n =?$), and chronic obstructive pulmonary disease ($n =?$). Imaging with hyperpolarized 3He was performed under an Institutional Review Board (IRB)-approved protocol with written informed consent obtained from each subject. In addition, all imaging was performed under a Food and Drug Administration approved physician's Investigational New Drug application (IND 57866) for hyperpolarized 3He. MRI data were acquired on a 1.5 T whole-body MRI scanner (Siemens Sonata, Siemens Medical Solutions, Malvern, PA) with broadband capabilities and a flexible 3He chest radiofrequency coil (RF; IGC Medical Advances, Milwaukee, WI; or Clinical MR Solutions, Brookfield, WI). During a 10–20-second breath-hold following the inhalation of $\approx$ 300 mL of hyperpolarized 3He mixed with $\approx$ 700 mL of nitrogen, a set of 19–28 contiguous axial sections were collected. Parameters of the fast low angle shot sequence for 3He MRI were as follows: repetition time msec / echo time msec, 7/3; flip angle, 10°; matrix, 80 × 128; field of view, 26 80 × 42 cm; section thickness, 10 mm; and intersection gap, none. The data were deidentified prior to analysis.

2.2 Algorithmic implementations

All algorithms and evaluation scripts are implemented within the ANTsR/ANTsRNet framework—a component of the ANTsX ecosystem (38) 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 (35) to improve performance and provide a more clinically granular labeling. In addition, further modifications incorporated additional data during training, added attention gating (39)to the U-net network (40), and

[2]Python versions are also available through ANTsPy/ANTsPyNet.

novel data augmentation strategies. More details are given below.

2.3.1 Network training

A 2-D per-image-slice U-net model (40) was trained with several parameters recommended by recent U-net exploratory work (41). 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 (42) 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`[3] and
- `histogramWarpImageIntensities`[4]

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

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" )
```

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

3 Results

3.1 T1-weighed brain segmentation analogy

As a preview of the

In Figure 3

Although the reference image set has been intensity normalized to $[0, 1]$ with truncated image intensities (quantiles = $[0, 0.99]$), it is apparent that the major features of the respective

image histograms (specifically, the three peaks which correspond to the cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM)) do not line up in this globally aligned space. Attempting to create a "reference" histogram from misaligned data is not without controversy. This can be seen in the results shown in the bottom where the linear binning analog drastically overstimates the amount of gray matter and simultaneously underestimates the amount of gray matter. The k-means approach, using precisely the same center clusters as determined via the reference histogram, yields a much better segmentation as it is optimizing the piecewise affine transform over histogram features. However, the hard threshold values result in labelings susceptible to noise in contrast to the GMM-MRF segmentation results.

3.2 Effect of reference image set selection

3.3 Effect of MR nonlinear intensity warping and additive noise

Need to add a SSIM calculation for each simulated image along with different histogram similarity measurements. We can then rescale all measurements for comparison and show how the SSIM calculation has lower variance than the histograms. THis shows that the image-to-histogram transformation results in information which is less robust than the original image.

3.4 Diagnostic prediction

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.

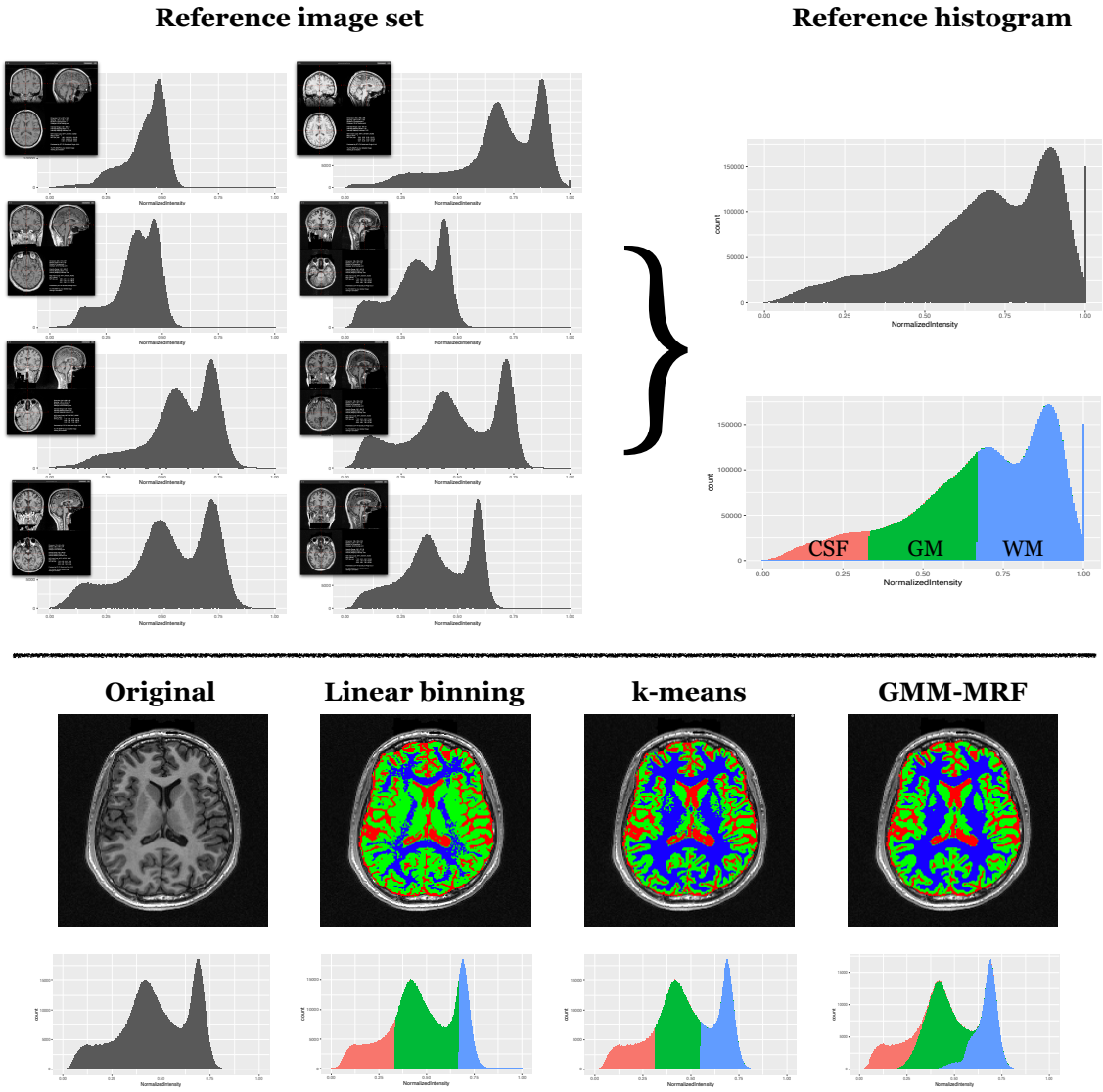
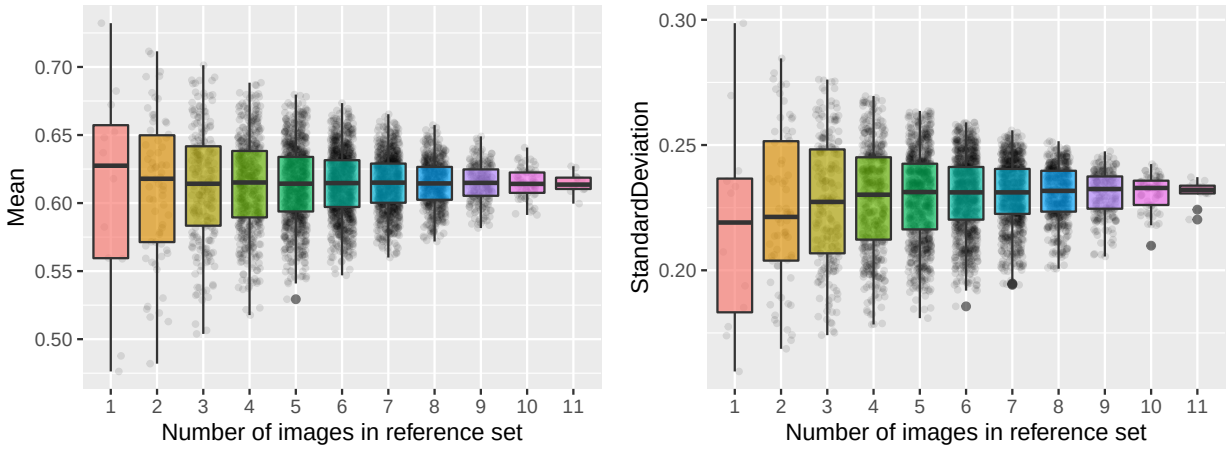


Figure 3: T1-weighted three-tissue brain segmentation analogy. Placing the three segmentation algorithms (i.e., linear binning, k-means, and GMM-MRF) in the context of brain tissue segmentation provides an alternative perspective for comparison. In the style of linear binning, we randomly select an image reference set using structurally normal individuals which is then used to create a reference histogram. (Bottom) For a subject to be processed, the resulting hard threshold values yield the linear binning segmentation solution as well as the initialization cluster values for both the k-means and GMM-MRF segmentations which are qualitatively different.

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.



(a) Mean reference plot.

(b) Standard deviation reference plot.

Figure 4

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.

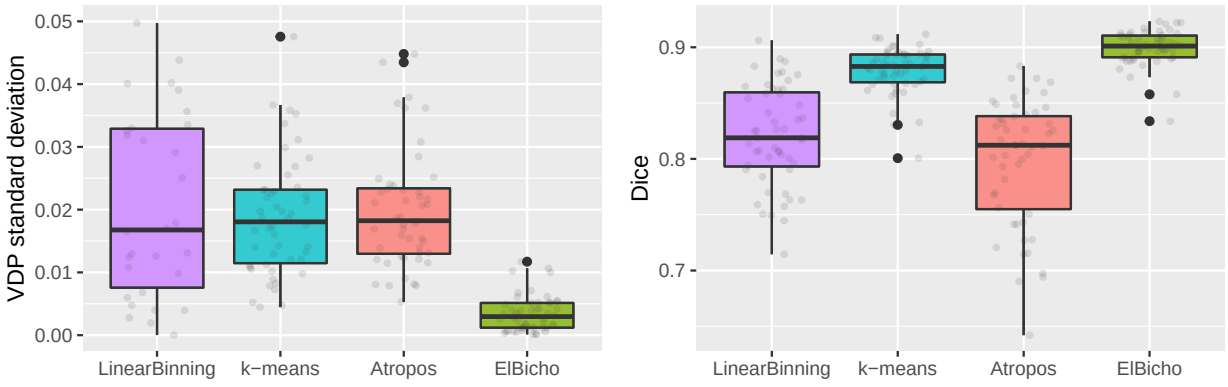
4. Altes TA, Powers PL, Knight-Scott J, et al.: Hyperpolarized 3He MR lung ventilation imaging in asthmatics: Preliminary findings. *J Magn Reson Imaging* 2001; 13:378–84.

5. Lange EE de, Mugler JP 3rd, Brookeman JR, et al.: Lung air spaces: MR imaging evaluation with hyperpolarized 3He gas. *Radiology* 1999; 210:851–7.

6. Samee S, Altes T, Powers P, et al.: Imaging the lungs in asthmatic patients by using hyperpolarized helium-3 magnetic resonance: Assessment of response to methacholine and exercise challenge. *J Allergy Clin Immunol* 2003; 111:1205–11.

7. 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.

8. Shammi UA, D'Alessandro MF, Altes T, et al.: Comparison of hyperpolarized 3He and



(a) Mean reference plot.

(b) Standard deviation reference plot.

Figure 5

129Xe mr imaging in cystic fibrosis patients. *Acad Radiol* 2021.

9. 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.

10. 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.

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

12. Kirby M, Svenningsen S, Owrangi A, et al.: Hyperpolarized 3He and 129Xe mr imaging in healthy volunteers and patients with chronic obstructive pulmonary disease. *Radiology* 2012; 265:600–10.

13. 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.

14. Thomen RP, Sheshadri A, Quirk JD, et al.: Regional ventilation changes in severe asthma after bronchial thermoplasty with (3)He mr imaging and ct. *Radiology* 2015; 274:250–9.

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

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

17. 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.

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

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

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

21. 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.

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

23. 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.

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

25. 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.

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

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

28. Besag J: On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society Series B (Methodological)* 1986; 48:259–302.

29. Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* 1977; 39:1–38.

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

31. Svenningsen S, McIntosh M, Ouriadov A, et al.: Reproducibility of hyperpolarized 129Xe mri ventilation defect percent in severe asthma to evaluate clinical trial feasibility. *Acad Radiol* 2020.

32. LeCun Y, Bengio Y, Hinton G: Deep learning. *Nature* 2015; 521:436–44.

33. Shen D, Wu G, Suk H-I: Deep learning in medical image analysis. *Annu Rev Biomed Eng* 2017; 19:221–248.

34. 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.

35. 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.

36. Warfield SK, Zou KH, Wells WM: Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004; 23:903–21.

37. Tustison NJ, Johnson HJ, Rohlfing T, et al.: Instrumentation bias in the use and evaluation of scientific software: Recommendations for reproducible practices in the computational

sciences. *Front Neurosci* 2013; 7:162.

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

39. 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.

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

41. 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.

42. 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.