Published in final edited form as:

Acad Radiol. 2019 July ; 26(7): 949–959. doi:10.1016/j.acra.2018.08.015.

A Comparison of Two Hyperpolarized ^{129}Xe MRI Ventilation Quantification Pipelines: The Effect of Signal to Noise Ratio

Mu He¹, Wei Zha², Fei Tan³, Leith Rankine⁴, Sean Fain^{2,5,6}, and Bastiaan Driehuys^{3,4,7}

¹Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States

²Department of Medical Physics, University of Wisconsin-Madison, Madison, WI, United States

³Department of Biomedical Engineering, Duke University, Durham, NC, United States

⁴Medical Physics Graduate Program, Duke University, Durham, NC, United States

⁵Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI, United States

⁶Department of Radiology, University of Wisconsin-Madison, Madison, WI, United States

⁷Department of Radiology, Duke University, Durham, NC, United States

Abstract

Rationale: Hyperpolarized ^{129}Xe MRI enables quantitative evaluation of regional ventilation. To this end, multiple classifiers have been proposed to determine ventilation defect percentage (VDP) as well as other cluster populations. However, consensus has not yet to be reached regarding which of these methods to deploy for multi-center clinical trials. Here, we compare two published classification techniques—linear-binning and adaptive *K-means*—to establish their limits of agreement and their robustness against reduced signal-to-noise ratio (SNR).

Methods: 29 subjects (age: 38.4 ± 19.0 years) were retrospectively identified for inter-method comparison. For each ^{129}Xe ventilation image, 7 reduced SNR image sets were generated with equal decrements relative to the native SNR. All 8 sets of images were then analyzed using both methods independently to classify all lung voxels into four clusters: VDP, Low-, Medium- and High-ventilation-percentage (LVP, MVP and HVP). For each cluster, the percentage of the lung it comprised was compared between the two methods, as well as how these values persisted as SNR was degraded.

Corresponding Author: Bastiaan Driehuys, Center for In Vivo Microscopy, Box 3302, Duke University Medical Center, Durham, NC 27710, Phone (919) 684-7786, Fax (919) 684-7158.

Authors' contributions: All authors designed the method in this study. All authors interpreted the data, prepared the manuscript, and drafted the article. All authors approved this manuscript in its final form.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest: MH, WZ, FT, LR, SF have no conflict of interest relevant to the study. BD is founder of Polarean, which is involved in the commercialization of hyperpolarized ^{129}Xe MRI technology.

Results: The limits of agreement for calculating VDP were $[+0.2\%, +4.0\%]$ with a $+1.5\%$ bias for binning relative to *K-means*. However, the inter-method agreement for the other clusters was moderate, with biases of -5.7% , 8.1% and -4.0% for LVP, MVP, and HVP respectively. As SNR decreased below ~ 4 , both methods began reporting values that deviated substantially from the native image. By requiring VDP to remain within $\pm 1.8\%$ of that calculated from the native image, the minimum tolerable SNR values were 2.4 ± 1.0 for the linear-binning, and 3.5 ± 1.5 for the *K-means*.

Conclusions: Both methods agree well in quantifying VDP, but agreement for LVP and MVP remains variable. We suggest a required SNR threshold be two standard deviations above the minimum value of 3.5 ± 1.5 required for robust determination of VDP, suggesting a minimum SNR of 6.6. However, robust quantification of the ventilated clusters required an SNR of 13.4.

Keywords

Hyperpolarized ^{129}Xe MRI; Quantification; Linear-binning; *K-means*

INTRODUCTION

Hyperpolarized (HP) gas MRI, using ^3He , and more recently ^{129}Xe , enables direct visualization of its breath-hold ventilation distribution in the lung (1–3). These images can be analyzed quantitatively to provide a sensitive means of monitoring regional disease and measuring therapy response (4). The first approaches used reader-based scoring that characterized ventilation defect numbers and sizes (5). Subsequently, Woodhouse et al (6), introduced a means to use the patient's own thoracic cavity ^1H MR image to calculate the percentage of ventilated lung volume. Later, Kirby et al (7), suggested instead, reporting the ventilation defect percentage (VDP), and proposed a *K-means* method of classifying ventilation into multiple clusters. Similarly, Tustison et al (8) also partitioned ventilation into multiple clusters, while further incorporating spatial contextual information. In recent years, VDP has emerged as among the most promising biomarkers in obstructive lung disease. It has been shown to correlate significantly with global pulmonary function test metrics (1, 2, 9, 10), and along with other biomarkers (11–14), has emerged as one of the most promising predictors of exacerbations in both COPD (9) and asthma (15). Moreover, it has been recognized that further insights may be gained by analyzing the remainder of the ventilation distribution. To this end, numerous quantification approaches have been developed to report metrics characterizing the ventilation distribution, such as heterogeneity, skewness and kurtosis, but also classification of the images into multiple intensity clusters (16–18). Although the importance of such quantitative analysis is undisputed, no consensus has yet emerged as to how to best determine VDP or other ventilation distribution parameters.

The challenge of quantifying ^{129}Xe MRI stems from its lack of an absolute intensity standard (such as Hounsfield units in CT). For a given patient, the ^{129}Xe intensity distribution depends not only on the underlying ventilation pattern, but also on the acquisition-specific factors such as ^{129}Xe polarization, inhaled volume, coil sensitivity and receiver gains. These must be accounted for in order to robustly and consistently extract quantitative metrics that can be compared over time and across patients. The majority of methods now take advantage of a companion breath-hold ^1H MRI scan to delineate the

thoracic cavity and confine the ^{129}Xe intensity analysis. However, this still leaves open decisions regarding how to parse the ^{129}Xe distribution to assign particular voxels to specific clusters. Without inter-method validation, it will be difficult to compare quantitative ventilation measures between centers and across patient populations.

Here, we seek to address these issues by comparing the performance of two quantification methods for ^{129}Xe ventilation MRI. Specifically, we compare the recently introduced histogram rescaling and binning approach (19, 20) with the more established *K-means* algorithm (7, 21, 22). Although several different variations of *K-means* analysis have been published, here we employ the most recent refinement of the *K-means* classifier, the adaptive *K-means* approach (21, 22).

Briefly, the binning approach assigns pixels in the ^{129}Xe ventilation scan to specific bins by rescaling the intensity histogram by its top percentile such that it ranges from 0–1. It then uses the standard deviation of a healthy reference population distribution (19), to set thresholds by which to assign pixels to four clusters referred to as: ventilation defect percentage (VDP), low-, medium, and high- ventilation percentage (LVP, MVP, HVP) respectively. The adaptive *K-means* instead employs an initial histogram analysis to determine whether disease is absent, mild-, or severe (21, 22), and based on this outcome employs two rounds of *K-means* clustering to determine the ventilation defect cluster, followed by three additional rounds of *K-means* to assign three additional clusters.

In comparing these methods, we sought to answer three questions. First – what are the limits of agreement between the methods for reporting the percentage occupancy of each cluster? Second – what is the sensitivity of the respective methods to low SNR conditions? Third – what are the implications of these findings for determining the required ^{129}Xe dose range to obtain adequate SNR? In doing so, we seek to help the pulmonary functional MRI community standardize its analysis methods for future multi-center studies. To facilitate this endeavor we have made the images, as well as their SNR-degraded versions, and thoracic cavity masks, publicly available at Harvard Dataverse (23).

METHODS

Subjects

The study employed a retrospective analysis of previously acquired, IRB-approved ^{129}Xe ventilation MRI scans (24). The dataset was comprised of 29 subjects, including 10 healthy controls (age: 25.7 ± 3.4 years, FEV1 %: $103.9 \pm 13.3\%$) and 19 patients (age: 45.1 ± 20.4 years, FEV1: 81.79 ± 19.3) with mild intermittent asthma.

MR Image Acquisition

All MR scans had been acquired on a 1.5 T GE Healthcare EXCITE 15M4 MR system, using protocols described previously (19). Briefly, subjects were scanned in the supine position with a flexible chest coil (Clinical MR Solutions, Brookfield, WI) that was tuned to the 17.66 MHz ^{129}Xe frequency and proton-blocked to permit acquiring anatomical scans using the ^1H body coil. All subjects underwent ^{129}Xe ventilation MRI after inhaling a dose-equivalent (DE) (25) of approximately 71 ± 11 ml HP ^{129}Xe filled to 1 L total volume with

buffer gas. Images were acquired in an anterior to posterior order using a fast spoiled gradient echo sequence with field of view (FOV) = 40 cm, matrix = $128 \times (90-128)$, slice thickness = 12.5 mm, bandwidth = 8.3 kHz, flip angle = $7^\circ-10^\circ$, and repetition time (TR)/echo time (TE) = 8.1/1.9 ms. A matching image of the thoracic cavity was acquired using the ^1H body coil prior to the ^{129}Xe scan. This allowed subjects to remain in the same position as for ^{129}Xe MRI and they inhaled a 1-L bag of room air to match lung inflation between the two acquisitions. The ^1H scan employed a steady state free precession imaging sequence with FOV = 40 cm, matrix = 192×192 , slice thickness = 12.5 mm, flip angle = 45° , TR/TE = 2.8/1.2 ms, and bandwidth = 125 kHz.

Degrading image SNR by adding noise to source images

The entire workflow is shown in Figure 1. The SNR for the source ^{129}Xe ventilation images for all subjects was 10.2 ± 5.1 . For each original ^{129}Xe ventilation image, 7 additional variants were generated, with progressively lower SNR down to ~1. Although the MR scans are generated from underlying complex data, with a Gaussian noise distribution (26–28), they are reconstructed as magnitude images, causing their noise to follow a Rician distribution (27–29). Moreover, prior to reconstruction on the scanner, a Fermi filter is typically applied to the raw k-space data, which affects the structure of the image-domain noise. Since only the reconstructed magnitude source images were available for this study, the following approach was used to generate SNR-degraded images with a realistic noise pattern. First, complex Gaussian noise with zero mean was generated in the spatial frequency domain and then Fermi filtered using parameters identical to those of the original image (radius 64, width 10 matrix units). These noise data sets were generated with a range of noise amplitudes such that the SNR of the noise-enhanced images ranged from the original SNR_0 down to 1, incrementing by ΔSNR according to

$$\Delta SNR = \frac{SNR_{original} - 1}{7} \quad [1]$$

Thus, generating images with individual SNR_i given by

$$SNR_i = SNR_{original} - \Delta SNR \times i \quad (i = 1, 2, \dots, 7) \quad [2]$$

In order to generate images with these desired SNR values, complex Gaussian noise was generated with a range of standard deviations σ_i that were calculated based on the SNR of native image as follows. For a magnitude image, the noise distribution will be Rician and its SNR given by

$$SNR = \frac{Mean_{signal} - Mean_{noise}}{Std_{noise}} \times \sqrt{2 - \frac{\pi}{2}} \quad [3]$$

where $Mean_{signal}$ is the average signal within the combined mask (thoracic cavity mask and ventilation mask), and $Mean_{noise}$ and Std_{noise} are the mean value and standard deviation of the background noise. (Note, the average signal within the combined mask includes regions of ventilation defects.) Following the properties of the Rician distribution (27), $Mean_{noise} \approx 1.25\sigma$ and $Std_{noise} \approx \sqrt{2 - \frac{\pi}{2}}\sigma$. Therefore, equation [3] can be simplified to derive the needed standard deviations of the seven target SNRs as follows:

$$\sigma_i = \frac{Mean_{signal}}{SNR_i + 1.25} \quad (i = 1, \dots, 7) \quad [4]$$

Seven noise-enhanced data sets were then generated with these standard deviations σ_i . After applying the Fermi filter, the noise sets were inverse Fourier transformed into the image domain and added to the original magnitude image to synthesize a noise-enhanced complex image. The resulting complex, noise-enhanced images were then added in quadrature to provide, SNR-degraded magnitude images for analysis.

Quantification workflow

All original ^{129}Xe ventilation images and noise-enhanced variants were processed through both the binning and adaptive *K-means* quantitative workflows. To limit the comparison to only the two intensity classifiers, several aspects of the workflow were first harmonized between the published methods.

Harmonization of methods prior to applying the classifier—The aspects of the two pipelines that were made identical between two algorithms are depicted in Figure 2. First, the ^1H thoracic cavity image was registered to the ^{129}Xe image (19). Second, the two were jointly segmented to obtain a combined binary mask that included both the thoracic cavity and trachea. Subsequently, the pulmonary vasculature was segmented out of the mask using a vesselness filter (8). Finally, the ^{129}Xe image intensity within the binary mask was corrected for B1 inhomogeneity using N4BiasCorrection (30) with default parameters with two exceptions. First, the algorithm was operated only on voxels within the combined ventilation and thoracic cavity mask. Second, the shrink factor that is used to accelerate the algorithm was reduced from its default value of 4 to 2 in order to provide robust results for gravity dependent signal variations when slices were acquired in the conventional anterior to posterior order as well as the opposite order.

Classification—The resulting intensity-corrected ^{129}Xe MRI data and lung masks were then fed into either the linear-binning (24) or the adaptive *K-means* (21) pipelines. Note, the binning method utilized the intensity voxels within a mask that included both the thoracic cavity and trachea to first rescale the histogram and classify the clusters, while the adaptive *K-means* method utilized only the thoracic cavity mask. However, both methods reported the percentages of classified voxels only for the volume within the thoracic cavity mask.

Both methods reported VDP, as well as the percentage of voxels populating the 3 higher signal bins labeled as LVP, MVP, and HVP. For linear-binning (Figure 3a), the B1-corrected,

re-scaled ^{129}Xe intensity histogram was classified into one of 6 bins by applying equally-spaced thresholds derived from a healthy reference population (24). However, to permit subsequent comparison with *K-means*-derived maps (containing only 4 clusters), bins 3–4 and 5–6 from the linear-binning maps were merged.

The adaptive *K-means* approach is illustrated in (Figure 3b). The intensity-corrected gas histogram of all lung pixels was first grouped into 10 equally spaced bins, the lowest of which determined the percentage of pixels P_L falling in the lowest-intensity bin. This value, established by a prior repeatability study (22), drove one of three pathways for calculating VDP and populating the ventilated clusters. Specifically, a first round of *K-means* was used to parse the data into 4 clusters if $P_L \geq 4\%$, and 5 clusters for $P_L < 4\%$. The lowest of these clusters, C1, then underwent a 2nd round of *K-means* to parse it into 4 sub-clusters, a combination of which formed the final VDP cluster depending on the value of P_L . For $P_L < 4\%$ (typical of healthy lungs), sub-clusters C1₁ and C1₂ comprised the VDP. For $4\% \leq P_L < 10\%$ (mild disease) C1₃ was added, and for $P_L \geq 10\%$ (severe-disease), all four sub-clusters of C1 comprised the VDP. After determining VDP, the remaining lung was classified into three ventilated clusters, using three rounds of adaptive *K-means*, again driven by the value P_L , as illustrated in Figure 3b. Further details of this approach are presented in the Appendix of reference (21).

Statistical Analysis

The Wilcoxon rank-sum test was used to determine whether VDP, LVP, MVP, and HVP derived from the two classifiers distinguished between control and asthma subjects. Bland-Altman plots were used to test for bias and to determine the limits of inter-method agreement on each global measure, while Dice coefficients were used to measure their level of regional agreement.

RESULTS

Agreement between Methods for Original ^{129}Xe MRI

Typical examples of ventilation maps generated by the two classification approaches from the original, high-SNR images are shown in Figure 4. In this example, depicting a healthy volunteer with few small defects, and an asthmatic subject with many modest-sized defects, both methods report qualitatively similar maps. A comparison of the cluster occupancies and overlap between the two pipelines across the entire cohort of asthma and control subjects is shown in Table 1. Both methods reported significantly greater VDP and lower MVP in asthma vs. controls ($p < 0.04$). However, in asthma, adaptive *K-means* found significantly higher LVP (the ventilated cluster immediately above VDP) relative to the normal group (Table 1). No significant difference was found for the HVP cluster classified by either binning or *K-means*. Comparing the spatial overlap of the two methods over the combined asthma and control data resulted in Dice coefficients of 0.4 ± 0.3 , 0.7 ± 0.1 , 0.9 ± 0.0 and 0.8 ± 0.1 for VDP, LVP, MVP, and HVP respectively.

Figure 5 shows the Bland-Altman plots comparing VDP, MVP, LVP and HVP calculated by the two classification methods. Only modest biases were found for VDP and HVP, with

linear-binning indicating slightly larger VDP (+1.5% bias) and slightly smaller HVP (−4.0% bias) with limits of agreement ranging from +0.2% to +4.0% for VDP and trending towards greater differences in asthma subjects, while ranging from −9.5% to 2.8% for HVP. However, larger biases were found for LVP (−5.7%) and MVP (+8.1%). The limits of agreement between the two methods were 0.5% to 13.4% for MVP, while LVP showed the broadest range from −11.7% to +1.9%.

Effects of SNR Degradation

For this cohort, the lowest SNR values were 1.2 ± 0.4 after applying evenly spaced noise degradation. An example of the effects of SNR degradation on the associated quantitative maps generated by the two methods is shown in Figure 6 for a subject with $VDP \approx 2\%$. In this case, both methods continue to report a fairly stable VDP until SNR falls below 4. A similar analysis is shown in Figure 7 for a subject exhibiting prominent ventilation defects on the original image. For this case, the methods diverge in how they are affected by decreasing SNR. Specifically, linear-binning reports a continuously decreasing VDP, which is caused by noise in the ventilation defect regions starting to be erroneously classified as low-ventilation signal. By comparison, for cases in which large ventilation defects are present in the source image, the adaptive *K-means* method appears more stable against the misclassification of defects and reports relatively stable VDP down to SNR ~ 1.8 .

The effects of SNR degradation across the entire cohort and for each of the clusters (VDP/LVP/MVP/HVP) are shown in Figure 8. Overall, the values for each cluster remained relatively stable for both methods as SNR decreased. In order to establish a minimum tolerable SNR for the quantification algorithms, we defined an empirical threshold for failure as the point at which the reported VDP deviated from the original VDP by $>1.8\%$; this threshold corresponds to the reported standard deviation of VDP within a healthy reference cohort (24). Using this tolerance threshold, the SNR at which the algorithms failed was significantly different between these two methods for the VDP ($p < 0.05$), LVP ($p < 0.05$) and HVP ($p < 0.05$) clusters, but not the MVP ($p = 0.82$) cluster. Specifically, the SNR at which the algorithms failed was 2.4 ± 1.0 for linear-binning and 3.5 ± 1.5 for adaptive *K-means* (Figure 8). Using a similar threshold of 1.8% for LVP, MVP and HVP, the lowest tolerable SNR values for LVP were 5.1 ± 2.6 for linear-binning, and 3.7 ± 1.8 for adaptive *K-means*, for MVP they were 5.6 ± 3.9 for linear-binning and 5.4 ± 2.5 for adaptive *K-means*, and for HVP they were 5.8 ± 3.8 for linear-binning and 3.8 ± 2.6 for adaptive *K-means*. Two methods report significantly different results for VDP

DISCUSSION

This study demonstrates a high degree of agreement between the linear-binning and adaptive *K-means* methods in classifying the ^{129}Xe MRI-derived VDP. Note that in this comparison all image pre-processing (thoracic cavity mask generation, registration, and bias field correction) were standardized so that only the classifiers were compared. Specifically, it was important to harmonize the approach to correcting the RF coil-induced B1 inhomogeneity that is common for the flexible coils used in pulmonary HP gas MRI (8, 19, 22). Specifically, we found that constraining the algorithm to operate within the combined lung

mask provided robustness against the level of disease severity and limiting the shrink factor to 2 ensured the N4BiasCorrection algorithm performed robustly regardless of slice acquisition order. Although the bias field correction was performed independently on the noisy data for all subjects, this work did not specifically investigate the effects of low SNR on the performance of the bias field algorithm. However, in the context of brain imaging, Tustison et al. (30) have confirmed its robustness against decreasing SNR.

While global agreement was strong between the methods in calculating VDP, they agreed to a significantly lesser extent in their classification of the other clusters (LVP, MVP, and HVP). This is perhaps not entirely surprising given the significant methodological differences. It is perhaps noteworthy that the adaptive *K-means* algorithm used here was trained using ^3He MRI data, which may possess somewhat different characteristics than ^{129}Xe MRI for the detection of ventilation defects (16). Fundamentally, both methods seek to characterize an intensity histogram with a scale that is initially arbitrary. Linear-binning treats this problem by rescaling the histogram using its top percentile and generating intensity bins by using a reference population. To normalize the signal intensities, it is important to note that the binning method retains signal from the trachea for this process on the assumption that the highest intensity signal has a high likelihood of originating from these voxels. *K-means* by contrast uses several rounds of refinement based on thresholds that were determined from a reproducibility study. However, *K-means* was therefore found to be highly sensitive to including trachea voxels, which typically dominate the high intensity cluster and lead to lower HVP (~2%) versus when the trachea is excluded (~15%). For this reason, the trachea was excluded from *K-means* analysis.

It is interesting to note that the Dice coefficients followed a different trend from the Bland-Altman analysis. Most notably while agreement on VDP was very strong, the Dice coefficient for this cluster was the lowest (0.4 vs 0.7 for all other clusters). This can be explained by the relatively small size of the VDP cluster, particularly for the healthy cohort, where the average value was 0.8% for binning and 1.8% for *K-means*. In this example, the methods clearly reported similar percentages, but with *K-means* reporting a cluster 2.25 times larger, the Dice coefficient by definition, could not exceed 0.6. Thus, interpretation of the Dice coefficient for such small clusters is limited.

When challenged with decreasing SNR, both methods continued to robustly report VDP until SNR had dropped to a relatively low value (3.5 ± 1.5 for *K-means*). The SNR requirements were somewhat higher for the other clusters, with calculation of HVP needing the highest value (5.8 ± 3.8 for linear-binning). It should be noted that determining the minimum acceptable SNR, necessitated setting an empirical threshold for failure relative to the values calculated for the original images. While this could be approached numerous ways, we used a tolerance of $\pm 1.8\%$, which represents our previously measured standard deviation in the value of VDP determined from a healthy reference cohort (24). This resulted in the minimum SNR values above and we propose to require an SNR two standard deviations higher than the minimum. Thus, for calculating VDP using *K-means*, this translates to an SNR of 6.6. Applying a similar approach to calculating HVP by linear-binning yields a required SNR of 13.4.

These SNR requirements can, in turn, be related to the minimum hyperpolarized ^{129}Xe dose required by using the previously established relationship between SNR, resolution, and dose (25). Briefly, the average SNR for fast, spoiled gradient echo imaging with typical ~8ms read-out times can be calculated according to:

$$\text{SNR} = 1.24 \text{ml}^{-2} \text{DE} \times V_{\text{vox}} \quad [5]$$

where SNR is the raw average SNR of the given dataset, V_{vox} is the voxel volume and DE is the dose equivalent given by:

$$\text{DE} = f_{129} \times P_{129} \times V_{\text{Xe}} \quad [6]$$

where f_{129} is the isotopic fraction of ^{129}Xe , P_{129} is its nuclear spin polarization and V_{Xe} is the xenon volume. Note the pre-factor of 1.24ml^{-2} is reduced by $\sqrt{2 - \frac{\pi}{2}}$ from our previously published scaling factor to properly account for the Rician nature of the background noise (28) in magnitude reconstructions in our study; see equation [3]. Using these principles for the highest requirement, that is SNR = 13.4 for the 0.12ml voxel volume used in this work, yields a required ^{129}Xe dose equivalent of 89.7 ml. Assuming 85% isotopic enrichment, such a dose can be achieved with 423 ml of xenon polarized to 25%.

This study shows that both methods report VDP values that are in close agreement. Moreover, both methods are remarkably robust against degrading SNR as well, with *K*-means perhaps providing a better estimate of VDP at low SNR. This may be attributable to the adaptive nature of the *K-means* thresholds, conferring perhaps a slight advantage against decreasing SNR. However, the two methods do not agree closely for the higher ventilation bins and, absent a gold standard, this makes determining a preferred method challenging. Perhaps one approach would be to test the reproducibility of the methods for analyzing test-retest images within the same patients. Alternatively, a choice could be made based on relative ease of interpretation. This work has revealed that while *K*-means methods can readily classify clusters above VDP, their interpretation may not be entirely straightforward. By contrast, it may be argued that the relatively simple approach of rescaling the intensity histogram is intuitive and is aided by the fact that its thresholds are derived from a healthy reference population, which is similar to typical medical test reporting. Agreement on these approaches to image quantification is critical to determining important factors such as the minimal clinically important difference for ^{129}Xe MRI VDP as was recently addressed for ^3He MRI (31).

Implications of harmonizing the ventilation quantification systems

While this work arguably represents a first step, achieving consensus around the image analysis approach is crucial to the advancement of this technology clinically. This scenario is not dissimilar from efforts to harmonize the analysis and quantification of positron emission tomography (PET) in order enable multicenter studies (32). Analysis software can be an important element of variability and defining standards for bias and precision is an essential

step to establishing a quantitative image biomarker as a measure of disease. Thus, much can be learned from this parallel field, which has strived to establish standards for acquisition/reconstruction, image quality, the choice of interpretation criteria and quantitative parameters, and the validity of the quantification methods. Paradigms for developing and validating quantitative image biomarkers such as the quantitative imaging network (QIN) project (33), and the quantitative imaging biomarkers alliance (QIBA) project (34), are supported by consortia of experts working together to define best approaches and standard methods while maintaining avenues for future innovation (35). Efforts such as the current work can begin to build a foundation for multicenter trials of ^{129}Xe MRI, and more specifically, as a first effort to assess the baseline reproducibility of existing analysis approaches. Once there is agreement on such standards for ventilation MRI, similar principles must be applied to other contrast mechanisms now being developed, such as imaging gas exchange (36, 37). In order to aid this conversation and allow for other methods to be compared, we have made the images, their SNR-degraded replicates, and lung masks used in this analysis publicly available (23).

Study Limitations

Although this initial comparison of the two quantification methods provides encouragement that harmonization will be feasible, this study did have some limitations. First, this retrospective study is necessarily limited as to the accuracy of the noise model used since only magnitude images were available for the reduced SNR comparisons. Second, estimating mean signal from the thoracic cavity in the presence of ventilation defects is debatable; we have opted not to exclude the defect regions from our mean calculation because the SNR calculated in this manner would remain the same for a given subject under inhalation of equal dose equivalent volumes of gas. Finally, the study used a relatively modest sample size and included only healthy subjects and those with asthma. While it did include a substantial range of defect severities and patterns, future work would benefit from prospectively recruiting a larger cohort, with a wider range of pulmonary disease conditions and severities. Moreover, such a study would also benefit from acquiring multiple scans on patients in order to rigorously assess repeatability; this could be achieved by expanding the availability of curated data sets gathered from multiple expert centers and archived as image examples for testing in future studies.

Acknowledgments

Funding: SARP grant (NIH/NHLBI - U10 HL109168), COAST grant (NIH/NHLBI - P01 HL070831), NIH/NHLBI - R01 HL126771. R01HL105643.

Reference:

1. de Lange EE, Altes TA, Patrie JT, et al. Evaluation of asthma with hyperpolarized helium-3 MRI - Correlation with clinical severity and spirometry. *Chest*. 2006; 130(4):1055–62. [PubMed: 17035438]
2. Mentore K, Froh DK, de Lange EE, Brookeman JR, Paget-Brown AO, Altes TA. Hyperpolarized HHe 3 MRI of the lung in cystic fibrosis: Assessment at baseline and after bronchodilator and airway clearance treatment. *Academic Radiology*. 2005; 12(11):1423–9. [PubMed: 16253854]
3. Mugler JP, 3rd, Altes TA. Hyperpolarized ^{129}Xe MRI of the human lung. *J Magn Reson Imaging*. 2013; 37(2):313–31. [PubMed: 23355432]

4. Svenningsen S, Eddy RL, Lim HF, Cox PG, Nair P, Parraga G. Sputum Eosinophilia and Magnetic Resonance Imaging Ventilation Heterogeneity in Severe Asthma. *Am J Respir Crit Care Med*. 2018; 197(7):876–84. [PubMed: 29313707]
5. 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(6):1205–11. [PubMed: 12789218]
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(4):365–9. [PubMed: 15779032]
7. Kirby M, Heydarian M, Svenningsen S, et al. Hyperpolarized He-3 Magnetic Resonance Functional Imaging Semiautomated Segmentation. *Academic Radiology*. 2012; 19(2):141–52. [PubMed: 22104288]
8. Tustison NJ, Avants BB, Flors L, et al. Ventilation-Based Segmentation of the Lungs Using Hyperpolarized He-3 MRI. *J Magn Reson Imaging*. 2011; 34(4):831–41. [PubMed: 21837781]
9. Kirby M, Pike D, Coxson HO, McCormack DG, Parraga G. Hyperpolarized He-3 Ventilation Defects Used to Predict Pulmonary Exacerbations in Mild to Moderate Chronic Obstructive Pulmonary Disease. *Radiology*. 2014; 273(3):887–96. [PubMed: 24960283]
10. Svenningsen S, Nair P, Guo FM, McCormack DG, Parraga G. Is ventilation heterogeneity related to asthma control? *Eur Respir J*. 2016; 48(2):370–9. [PubMed: 27174885]
11. Siva R, Green RH, Brightling CE, et al. Eosinophilic airway inflammation and exacerbations of COPD: a randomised controlled trial. *Eur Respir J*. 2007; 29(5):906–13. [PubMed: 17301099]
12. Han MLK, Kazerooni EA, Lynch DA, et al. Chronic Obstructive Pulmonary Disease Exacerbations in the COPD Gene Study: Associated Radiologic Phenotypes. *Radiology*. 2011; 261(1):274–82. [PubMed: 21788524]
13. Keene JD, Jacobson S, Kechris K, et al. Biomarkers Predictive of Exacerbations in the SPIROMICS and COPD Gene Cohorts. *Am J Resp Crit Care*. 2017; 195(4):473–81.
14. Petsky HL, Li A, Chang AB. Tailored interventions based on sputum eosinophils versus clinical symptoms for asthma in children and adults. *Cochrane Db Syst Rev*. 2017; (8).
15. Mummy DG, Kruger SJ, Zha W, et al. Ventilation defect percent in helium-3 magnetic resonance imaging as a biomarker of severe outcomes in asthma. *J Allergy Clin Immunol*. 2018; 141(3): 1140–1 e4. [PubMed: 29129582]
16. Svenningsen S, Kirby M, Starr D, et al. Hyperpolarized He-3 and Xe-129 MRI: Differences in Asthma Before Bronchodilation. *J Magn Reson Imaging*. 2013; 38(6):1521–30. [PubMed: 23589465]
17. Tzeng YS, Lutchen K, Albert M. The difference in ventilation heterogeneity between asthmatic and healthy subjects quantified using hyperpolarized He-3 MRI. *Journal of Applied Physiology*. 2009; 106(3):813–22. [PubMed: 19023025]
18. Virgincar RS, Cleveland ZI, Kaushik SS, et al. Quantitative analysis of hyperpolarized 129Xe ventilation imaging in healthy volunteers and subjects with chronic obstructive pulmonary disease. *NMR Biomed*. 2013; 26(4):424–35. [PubMed: 23065808]
19. He M, Kaushik SS, Robertson SH, et al. Extending Semiautomatic Ventilation Defect Analysis for Hyperpolarized Xe-129 Ventilation MRI. *Academic Radiology*. 2014; 21(12):1530–41. [PubMed: 25262951]
20. He M, Robertson SH, Wang JM, Rackley CR, McAdams HP, Driehuys B. Differentiating Early Stage And Later Stage IpF Using Hyperpolarized 129xe Ventilation Mri. *Am J Resp Crit Care*. 2016; 193.
21. Zha W, Kruger SJ, Cadman RV, et al. Regional Heterogeneity of Lobar Ventilation in Asthma Using Hyperpolarized Helium-3 MRI. *Acad Radiol*. 2018; 25(2):169–78. [PubMed: 29174189]
22. Zha W, Niles DJ, Kruger SJ, et al. Semiautomated Ventilation Defect Quantification in Exercise-induced Bronchoconstriction Using Hyperpolarized Helium-3 Magnetic Resonance Imaging: A Repeatability Study. *Acad Radiol*. 2016; 23(9):1104–14. [PubMed: 27263987]
23. He M, Zha W, Tan F, Rankine L, Fain S, Driehuys B. SNR-degraded 129Xe ventilation MRI for the comparison of quantification methods. *Harvard Dataverse*; 2018.

24. He M, Driehuys B, Que L, Huang Y-CT. Using Hyperpolarized ^{129}Xe MRI to Quantify the Pulmonary Ventilation Distribution. *Acad Radiol*. 2016.
25. He M, Robertson SH, Kaushik SS, et al. Dose and pulse sequence considerations for hyperpolarized (^{129}Xe) ventilation MRI. *Magn Reson Imaging*. 2015; 33(7):877–85. [PubMed: 25936684]
26. Coupe P, Manjon JV, Gedamu E, Arnold D, Robles M, Collins DL. Robust Rician noise estimation for MR images. *Med Image Anal*. 2010; 14(4):483–93. [PubMed: 20417148]
27. Dietrich O, Raya JG, Reeder SB, Ingrisch M, Reiser MF, Schoenberg SO. Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics. *Magn Reson Imaging*. 2008; 26(6):754–62. [PubMed: 18440746]
28. Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. *Magn Reson Med*. 1995; 34(6):910–4. [PubMed: 8598820]
29. Henkelman RM. Measurement of Signal Intensities in the Presence of Noise in Mr Images. *Med Phys*. 1985; 12(2):232–3. [PubMed: 4000083]
30. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: Improved N3 Bias Correction. *Ieee T Med Imaging*. 2010; 29(6):1310–20.
31. Eddy RL, Senningsen S, McCormack DG, Parraga G. What is the Minimal Clinically Important Difference for ^3He MRI Ventilation Defects? *Eur Respir J*. 2018.
32. Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *Eur J Nucl Med Mol Imaging*. 2017; 44(Suppl 1):17–31.
33. Committee QIN. Quantitative Imaging Network. Available at: https://imaging.cancer.gov/programs_resources/specialized_initiatives/qin/about/default.htm.
34. Committee QIBA. Quantitative Imaging Biomarkers Alliance Available at: http://qibawiki.rsna.org/index.php/Main_Page.
35. Committee QLDB. Harmonization CT Density Measures Across Platforms. Available at: https://qibawiki.rsna.org/images/7/72/LUNG-DENSITY-Poster_QIBA_Kiosk_RSNA2016.pdf.
36. Wang Z, Robertson SH, Wang J, et al. Quantitative analysis of hyperpolarized (^{129}Xe) gas transfer MRI. *Med Phys*. 2017; 44(6):2415–28. [PubMed: 28382694]
37. Wang Z, He M, Bier EA, et al. Hyperpolarized ^{129}Xe gas transfer MRI: the transition from 1.5T to 3T. *Magn Reson Med*. 2018.

Figure 1.

SNR-degradation protocol. a) Example of complex Gaussian noise generated in the frequency domain. b) Application of a Fermi filter with radius = $64/128$, width = $10/128$ to match the reconstruction parameters. c) Real component of the noise after 2D inverse Fourier transformation to the image space. d) Original hyperpolarized ^{129}Xe ventilation magnitude image. e) Magnitude of the ^{129}Xe ventilation image with imposed noise.

Figure 2.

Registration, masking, and bias field correction were shared between pipelines: (a) ^{129}Xe MRI as acquired. (b) ^1H thoracic cavity image registered to ^{129}Xe MRI. (c) Joint segmentation of the ^{129}Xe and ^1H images to obtain an initial mask for bias field (B1 inhomogeneity) correction. (d) ^{129}Xe MRI after bias-field correction. (e) Detection of vascular structures from ^1H MRI. (f) Refined thoracic cavity mask with vascular structures removed.

Figure 3.

Linear-binning and adaptive *K-means* algorithms. The binning method rescales the (Aa) ^{129}Xe intensities within the thoracic cavity mask by its top percentile such that it ranges from 0–1. (Ab) Each pixel is then assigned to an intensity cluster using thresholds based on the standard deviation of a healthy reference population. (Ac) Resulting linear-binning map. (Bb.(1)) The adaptive *K-means* decimates the original histogram of the lungs to 10 clusters. It then tests to see whether the first cluster (lowest intensity percentage P_L) contains <4%, 4–10% or 10% of all lung voxels. Depending on the P_L value, the defect cluster was determined after two rounds of *K-means*. (Bb.(2)) Setting aside the defect voxels, low-, medium-, and high-ventilation clusters were further classified from the ventilated voxels based on P_L after another two rounds of *K-means*. (Bc) Resulting adaptive *K-means* map.

Figure 4.

Comparison of ventilation maps determined by the linear-binning and adaptive *K-means* methods for both a healthy control and an asthma subject. Qualitatively, the two methods appear to report similar VDP, LVP, MVP and HVP clusters in both subjects.

Figure 5.

Bland-Altman plots comparing the *K-means* and binning analyses of the original high-SNR scans for 29 subjects. One plot is dedicated to each class of ventilation distribution — ventilation defect percentage (VDP), low ventilation percentage (LVP), medium ventilation percentage (MVP), and high ventilation percentage (HVP). These results show only modest bias between binning and adaptive *K-means* for VDP (+1.5%) and HVP (−4.0%), but more substantial biases for LVP (−5.7%) and MVP (+8.1%). The limits of agreement between the two methods were 0.5% to 13.4% for MVP, while LVP showed the broadest range from −11.7% to +1.9%.

Figure 6.

Representative ^{129}Xe ventilation MRI and associated linear-binning and adaptive *K-means* maps for a young asthmatic subject with low VDP as SNR is progressively degraded. Both methods appear to report a stable VDP as SNR decreases down to ~4. At that point, both methods start to report a substantially higher VDP than that derived from the un-degraded image.

Figure 7. Representative ^{129}Xe ventilation MRI and associated linear-binning and adaptive *K-means* maps for an older asthmatic subject with high VDP as SNR is progressively degraded. As SNR decreases, linear-binning reports a continuously decreasing VDP caused by noise in the ventilation defect regions being erroneously classified as low-ventilation signal. The adaptive *K-means* method appears more stable against such misclassification and reports relatively stable VDP down to SNR ~ 1.8 .

Figure 8.

Testing classification fidelity for linear-binning vs. adaptive *K-means* when challenged with decreasing SNR. Subjects with blue marks are analyzed using the linear –binning method, while subjects with red marks are analyzed using the adaptive *K-means* method. The empirical failure threshold was set at 1.8% difference in VDP (dotted lines) from the values obtained from the original image. With this threshold, the SNR at which the VDP value deviated from tolerance was 2.4 ± 1.0 for linear-binning and 3.5 ± 1.5 for adaptive *K-means*. Using a similar threshold for LVP, MVP and HVP, the lowest tolerable SNR values for LVP were 5.1 ± 2.6 for linear-binning, and 3.7 ± 1.8 for adaptive *K-means*, for MVP they were 5.6 ± 3.9 for linear-binning and 5.4 ± 2.5 adaptive *K-means*, and for HVP they were 5.8 ± 3.8 for linear-binning and 3.8 ± 2.6 adaptive *K-means*.

Table 1.

Ventilation distribution (VDP, LVP, MVP and HVP) reported by linear binning and adaptive K-means along with Dice coefficients reflecting their overlap. Both methods reported significantly higher ventilation defect percentage (VDP) and lower medium ventilation percentage (MVP) in asthma vs. control.

(%)	Linear Binning		Adaptive <i>K-means</i>		Dice Coefficient		
	Control (N=10)	Asthma (N=19)	Control (N=10)	Asthma (N=19)	Control (N=10)	Asthma (N=19)	All subjects (N=29)
VDP	0.8±1.4	6.1±9.4*	1.8±1.6	7.9±9.7*	0.3±0.2	0.5±0.2	0.4±0.3
LVP	16.5±3.7	20.3±6.7	9.0±2.8	15.6±6.7*	0.7±0.1	0.8±0.1	0.7±0.1
MVP	65.1 ±4.7	58.2±11.2*	76.3±4.6	64.6±12.7*	0.9±0.0	0.9±0.0	0.9±0.0
HVP	17.6±2.2	15.5±5.9	12.9±2.4	11.9±5.3	0.9±0.1	0.8±0.1	0.8±0.1

Values are expressed as mean ± standard deviation. The symbol * denoted significant difference ($p<0.05$) between asthma vs. control was observed in a ventilation level measured